CodeScout: An Effective Recipe for Reinforcement Learning of Code Search Agents

Lintang Sutawika*
lsutawik@cs.cmu.edu
Carnegie Mellon University

Aditya Bharat Soni*
adityabs@cs.cmu.edu
Carnegie Mellon University

Bharath Sriraam R R
Independent

Apurva Gandhi
Carnegie Mellon University

Taha Yassine
Independent

Sanidhya Vijayvargiya
Carnegie Mellon University

Yuchen Li
Independent

Xuhui Zhou
Carnegie Mellon University

Yilin Zhang
Carnegie Mellon University

Leander Melroy Maben
Carnegie Mellon University

Graham Neubig
gneubig@cs.cmu.edu
Carnegie Mellon University,
OpenHands

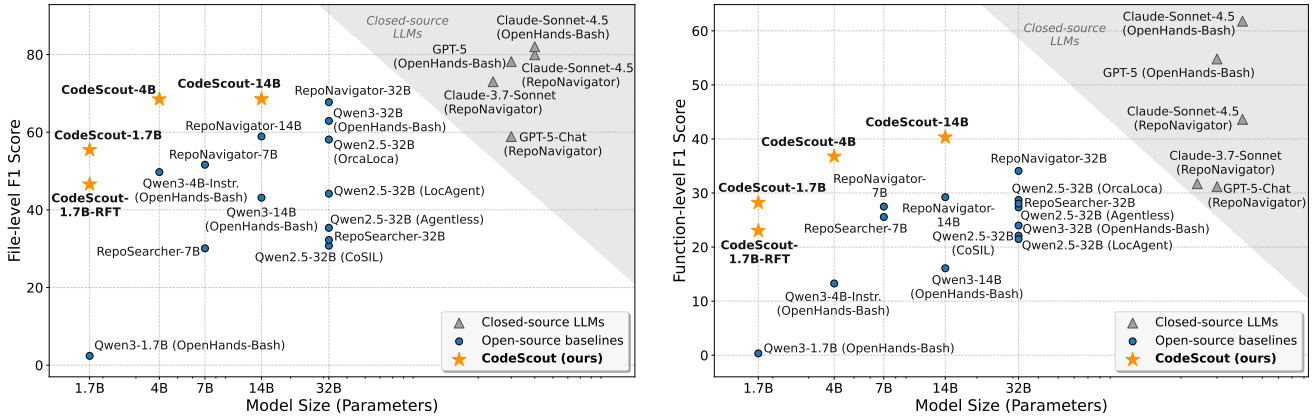


Figure 1: An overview of code localization performance of various approaches on SWE-Bench Verified. CodeScout achieves superior or competitive results over larger SoTA open-source LLMs and closes the gap with frontier closed-source LLMs.

Abstract

A prerequisite for coding agents to perform tasks on large repositories is code localization - the identification of relevant files, classes, and functions to work on. While repository-level code localization has been performed using embedding-based retrieval approaches such as vector search, recent work has focused on developing agents to localize relevant code either as a standalone precursor to or interleaved with performing actual work. Most prior methods on agentic code search equip the agent with complex, specialized tools, such as repository graphs derived using static analysis. In this paper, we demonstrate that, with an effective reinforcement learning recipe, a coding agent equipped with *nothing more* than a standard Unix terminal can be trained to achieve strong results. Our experiments on three benchmarks (SWE-Bench Verified, Pro, and Lite) reveal that our models consistently achieve superior or competitive performance over **2-18**× larger base and post-trained LLMs and sometimes approach performance provided by closed models like Claude Sonnet, even when using specialized scaffolds. Our work particularly focuses on techniques for re-purposing existing coding agent environments for code search, reward design, and RL optimization. We release the resulting model family, CodeScout, along with all our code and data for the community to build upon.

*Both authors contributed equally to this research.

ACM CAIS Workshop on Agentic SE '26, San Jose, CA

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Computing methodologies** → **Reinforcement learning**; *Intelligent agents*; • **Software and its engineering** → **Software maintenance tools**; **Software testing and debugging**.

Keywords

LLM agents, RL, Code search, Coding agent, Retrieval, Reinforcement Learning

ACM Reference Format:
Lintang Sutawika, Aditya Bharat Soni, Bharath Sriraam R R, Apurva Gandhi, Taha Yassine, Sanidhya Vijayvargiya, Yuchen Li, Xuhui Zhou, Yilin Zhang, Leander Melroy Maben, and Graham Neubig. 2026. CodeScout: An Effective Recipe for Reinforcement Learning of Code Search Agents. In *Proceedings of (ACM CAIS Workshop on Agentic SE '26).* ACM, New York, NY, USA, 35 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

For repository-level coding tasks such as those in the popular SWE-Bench benchmark [15], a critical first step is **code localization**: given an issue description and a codebase, the system must identify the relevant files and finer-grained code entities (e.g., classes and functions) to modify (§2; [12, 34]). This is challenging in large repositories with complex inter-dependencies, and relying on general-purpose language models to solve this localization problem as part of the agentic coding loop can result in high costs, incorrect fixes, and code bloat [11].

To address this issue, it is also common to incorporate some variety of specialized localization module to acquire the relevant context from large codebases in a more efficient and effective manner. Traditionally, this code search was performed through semantic code search using vector databases [33, 34]. In contrast, recent methods have investigated *agentic code search* - using agents to iteratively navigate the repository and uncover necessary evidence for solving the issue under consideration. Typically, these methods have involved significant modifications to the agent itself to incorporate static analysis of the codebase, such as LocAgent's repository graph navigation [3] and RepoNavigator's "jump" tool that retrieves definitions of Python symbols [42]. While well-grounded, these necessitate the use of static analysis tools tailored to a *particular* programming language, increasing operational complexity of deploying such agents on a broader variety of coding scenarios. Simultaneously, there have been various anecdotal reports from industry regarding reinforcement learning methods that increase the ability of agents to perform localization rapidly [5, 22], with varying levels of detail but without a clear recipe. Given this background, we address the fundamental research question: *given an appropriate reinforcement learning algorithm, what is an effective recipe to train a code localization agent that simply uses the terminal tool typical of more generic coding agents, yet achieves competitive or superior accuracy over agents using special-purpose tools?* This work provides the first demonstration of such a recipe, that involves scalable methods for data and environment creation (§3.1), a standard agent scaffold (§3.2), careful attention to reward design (§3.3) and training algorithm (§3.4). Our experiments (§4, §5) show that the resulting model, CodeScout, achieves state-of-the-art results on three localization datasets derived from SWE-Bench Verified, Pro, and Lite. Notably, all models and code are released publicly for future research to build upon.[1]

Table 1: Comparison of repository-level code localization methods. CodeScout is the only approach that directly post-trains LLMs with RL using a simple, programming language-agnostic agent scaffold equipped *solely* with a bash terminal.

Method	Language-Agnostic Scaffold	Pure RL post-training	# Tools
LocAgent [3]	✗	✗	3
CoSiL [14]	✗	✗	3
OrcaLoca [39]	✗	✗	5
RepoSearcher [20]	✗	✗	5
RepoNavigator [42]	✗	✓	1
CodeScout *(Ours)*	✓	✓	1

2 Related Work

Prior work proposes specialized agents and trains open-source LLMs for code localization (Table 1), but is often limited to a single programming language (typically Python) due to reliance on static analysis tools (e.g., AST parsers). Extending them to other languages requires additional engineering effort.

LocAgent [3] and OrcaLoca [39] construct code graphs to capture hierarchical dependencies, requiring expensive pre-indexing. CoSIL [14] dynamically builds module and function call graphs, and RepoSearcher [20] has specialized tools, for example to extract imports or search for methods in a class. While CoSIL and RepoSearcher avoid pre-indexing, they still require static analysis. RepoNavigator [42] has a "jump" tool that resolves Python symbol definitions using a language server built using AST. Our agent *solely* uses a terminal, making it inherently language-agnostic by design.

Most prior methods either do not train LLMs using their proposed agents, like CoSIL and OrcaLoca, or rely on supervised distillation from closed-source LLMs via rejection sampling fine-tuning [40], like LocAgent and RepoSearcher. CodeScout trains LLMs directly with RL, removing dependence on expensive proprietary LLMs for data curation.

Finally, CodeScout has a *significantly simpler* scaffold both in terms of both engineering overhead and tool count. Since bash terminal is already a core component of standard coding agents [1, 32, 37], we do not need to implement specialized task-specific tools or scaffolds, unlike prior work. Moreover, when comparing the size of the agent's action space, excluding any finish tools, CodeScout has only 1 tool, compared to 3-5 tools in prior work. Appendix D includes a detailed review of prior work.

3 CodeScout: An Effective RL Recipe for Code Localization

This section presents the methodology used to train CodeScout, covering training data curation and environment construction (§3.1), the agent scaffold (§3.2), reward design (§3.3), and the RL training setup (§3.4). Figure 2 provides an overview of our approach.

[1] Code: https://github.com/OpenHands/codescout. HuggingFace Collection: https://huggingface.co/collections/OpenHands/codescout

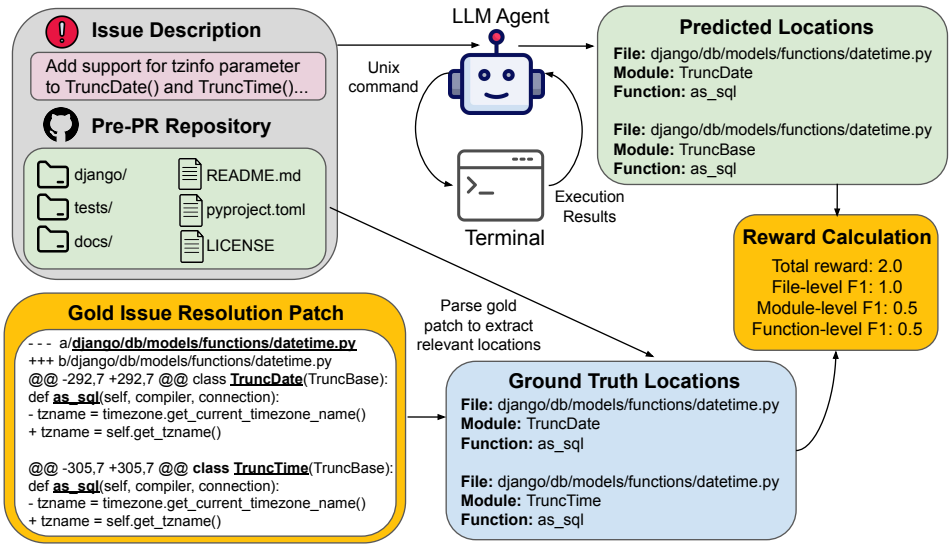


Figure 2: An overview of CodeScout: given an issue, the agent navigates the pre-PR codebase using a terminal and predicts the relevant files, modules, and functions. The reward function computes F1 scores for these three granularities using ground truth locations extracted from the gold issue resolution patch.

3.1 Data and Environment Curation

Given a GitHub issue $\mathcal{I}$ in a Python repository $\mathcal{R}$, we process the ground-truth issue resolution patch $\mathcal{P}$ to extract the localization targets at three granularities. Specifically, we define the ground-truth target as $y^\star = (F^\star, M^\star, U^\star)$, where $F^\star = f_1^\star, \ldots, f_{N_f}^\star$ is the set of modified files, $M^\star = m_1^\star, \ldots, m_{N_m}^\star$ is the set of modified modules, and $U^\star = u_1^\star, \ldots, u_{N_u}^\star$ is the set of functions or methods edited by the patch $\mathcal{P}$. An example of these targets is illustrated in Figure 2. The LLM agent is tasked with predicting $y^\star$ given the issue description $\mathcal{I}$ and the pre-PR repository state $\mathcal{R}$. We extract ground truth by using patch-processing scripts from Chen et al. [3].

We curate training instances by processing GitHub issues collected by prior work [23, 38], originally intended for training agents to fix issues. We discard issues with empty descriptions and whose PRs create or delete files, as the agent cannot predict the name of newly created files and we cannot determine relevant modules or functions for deleted files. We ignore non-Python files (e.g., `README.md`) because function and module-level information cannot be extracted from them. We construct the RL environment by cloning the pre-PR commit of the repository to a specific path known to the agent via its prompt. As localization does not require code execution, we do not install dependencies or use sandboxes. Notably, as our agent only uses a terminal, its environment setup overhead is much lower than prior agents that use code graphs, vector indices, or dependency parsers.

3.2 OpenHands-Bash: Our Agent Scaffold

We use OpenHands Agent SDK [32] to implement our agent as it performs strongly across software engineering benchmarks, implements core tools (e.g., terminal), supports parallel tool-calling, and has a modular design that easily integrates with our training backend. The agent is equipped with a `Terminal` tool that supports standard Unix commands (e.g., `find`, `ls`, `sed`). We install `ripgrep` [8] in the agent's environment, which is a command-line utility for fast grep-style search using regex patterns recursively through all code files in a directory. Furthermore, the agent uses the `LocalizationFinish` tool to submit predictions. We refer to this scaffold as **OpenHands-Bash**.

Our initial experiments used a string-based output format from Chen et al. [3], but we found training with this format is sensitive to noisy reward signals due to brittle validation and parsing. We address this by requiring the agent to terminate via the `LocalizationFinish` tool, enforcing a structured output schema and simplifying parsing, improving reward fidelity. Appendix A includes detailed prompts and tool definitions for our agent.

3.3 Reward Design

For each trajectory τ sampled from the agent, we extract the prediction $y = (F, M, G)$ from the `LocalizationFinish` tool, which specifies the predicted files, modules, and functions. Given the ground-truth $y^\star = (F^\star, M^\star, G^\star)$, we compute F1 scores at each granularity. For each set $S \in \{F, M, G\}$ with the corresponding ground truth set $S^\star \in \{F^\star, M^\star, G^\star\}$, we compute F1 score (i.e. harmonic mean of precision and recall): $r^{\text{F1-file}}$, $r^{\text{F1-module}}$, and $r^{\text{F1-func}}$. Our reward function is defined as:

$$r(\tau, y, y^\star) = r^{\text{F1-file}}(y, y^\star) + r^{\text{F1-module}}(y, y^\star) + r^{\text{F1-func}}(y, y^\star) \tag{1}$$

Earlier training runs for our 14B LLM collapse to near-zero rewards in later stages as the agent frequently exhausted the step budget without submitting its answer. We fix this with an auxiliary binary reward $r^{\text{turn}}(\tau, k)$ that assigns 1 if and only if the agent terminates in exactly k turns, where k is the step limit. Thus, the reward for 14B

LLM adds this binary term to $r(\tau, y, y^\star)$ in Equation 1. Appendix G includes more discussion on reward design.

3.4 RL Training Algorithm

We use SkyRL [10] and train asynchronously [7] to improve GPU utilization by parallelizing rollouts and weight optimization and set maximum staleness to $t = 4$. We train models using Group Sequence Policy Optimization (GSPO) [44]. Given model parameters θ ; i indexes sequences in a group of size G; y_i is the i^{th} output sequence with length $|y_i|$, and $y_{i,t}$ is its t^{th} token. π_θ and $\pi_{\theta_{\text{old}}}$ are current and older policies, $\hat{A}_i$ is the advantage, and ε is the clipping parameter:

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_i \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(s_i(\theta)\hat{A}_i,\ \text{clip}\left(s_i(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_i \right) \right] \quad (2)$$

where, s_i is the importance ratio derived from sequence likelihood [43]:

$$s_i(\theta) = \exp\left(\frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \log \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})} \right) \quad (3)$$

Following Liu et al. [17], we remove standard deviation from the advantage calculation.

$$\hat{A}_i = r_i - \text{mean}(\mathbf{r}), \mathbf{r} = \{r_1, ..., r_G\} \quad (4)$$

Finally, we remove KL regularization from the loss, disable entropy loss, and mask loss for rollouts that exhaust maximum steps without submitting predictions. Additional analysis on the effect of the RL algorithm on localization performance can be found in Appendix H.

4 Experimental Setup

We discuss CodeScout's experimental setup for training (§4.1) and evaluation (§4.2).

4.1 Training Setup

We process the SWE-Smith [38] dataset using the method from §3.1, resulting in a training set of 39K instances from 128 repositories. None of these repositories are present in our evaluation benchmarks (§4.2), avoiding any risk of dataset contamination. We train LLMs from Qwen3 series [36]: `Qwen3-1.7B`, `Qwen3-4B-Instruct-2507`, and `Qwen3-14B`. We disable thinking for 1.7B and 14B LLMs and mask loss for tokens not generated by the LLM like system and user prompts, and tool execution results.

CodeScout-4B and CodeScout-14B are directly trained from their corresponding base models using our modified GSPO algorithm (§3.4). As the `Qwen3-1.7B` base LLM performs very poorly (§5), we do not directly train it using RL. Instead, we warm-start the model with rejection sampling fine-tuning (RFT) [40] on 4K successful trajectories (i.e., F1 = 1.0 for file, module, and function; see §4.2) sampled from CodeScout-14B. The resulting model (CodeScout-1.7B-RFT) is further trained using RL (§3.4) on samples not seen during RFT resulting in CodeScout-1.7B. Appendix B.1 provides additional experimental details.

4.2 Evaluation Setup

We report results on SWE-Bench Verified [4] (500 instances), SWE-Bench Lite [15] (300 instances), and the Python subset of SWE-Bench Pro [6] (266 instances), which is more challenging than the other two. We re-purpose the datasets to extract ground-truth locations (§3.1). Our primary evaluation metric is the instance-wise average of F1 score between prediction and ground-truth for three granularities: file, module, and function. While we also report average precision and recall for each granularity, we mainly focus on F1 score as it captures both precision and recall.

Baselines: We compare CodeScout against several baselines using both open-source and closed-source LLMs: RepoNavigator [42], RepoSearcher [20], LocAgent [3], OrcaLoca [39], CoSIL [14], and Agentless [34]. All baseline metrics are taken from prior work. While Chen et al. [3], Jiang et al. [14], Ma et al. [20], Xia et al. [34] output a ranked list of top-K locations (typically K is fixed and set to 5), Yu et al. [39], Zhang et al. [42] and CodeScout dynamically predict a variable number of locations, which better reflects the fact that issues may require editing varying number of locations. Thus, some baselines have severe precision-recall disparities. Appendix J includes additional discussion.

We evaluate other LLMs with OpenHands-Bash (§3.2): the base variants of CodeScout (`Qwen3-4B-Instruct-2507`, and the non-thinking versions of `Qwen3-1.7B` and `Qwen3-14B`), and `Qwen3-32B` with thinking enabled to compare with a reasoning LLM. We also benchmark GPT-5 and Claude-Sonnet-4.5. Despite specifying the turn limit in the system prompt, GPT-5 and Claude-Sonnet-4.5 frequently exhaust all steps without submitting answers. Following Zhang et al. [42], we fix this by adding a reminder before the last turn that prompts the LLM to submit its final answer. Appendix B.2 includes more details on our evaluation setup.

5 Results

This section describes our experimental results. Table 2 summarizes results on all three benchmarks, and we include detailed results for SWE-Bench Verified in Table 3, SWE-Bench Pro in Table 4, and SWE-Bench Lite in Table 5. We describe the key takeaways below.

5.1 CodeScout outperforms 8-18x larger base LLMs with OpenHands-Bash

For all benchmarks and search granularities, CodeScout LLMs *significantly* outperform their corresponding base models with the OpenHands-Bash agent. CodeScout-1.7B achieves absolute gains in F1 score over its base LLM of **40-54%** for files, **25-41%** for modules, and **18-28%** for functions. CodeScout-4B yields absolute gains of **14-19%**, **25-38%**, and **21-31%** in file, module, and function-level F1 scores respectively. CodeScout-14B has absolute F1 gains of **23-34%** for files, **25-39%** for modules, and **20-33%** for functions.

CodeScout LLMs demonstrate exceptional parameter efficiency and consistently outperform significantly larger base LLMs with OpenHands-Bash. CodeScout-1.7B outperforms the **8**× larger Qwen3-14B with absolute F1 gains of **11-18%** for files, **13-21%** for modules, and **10-15%** for functions. When compared against the **18**× larger Qwen3-32B, CodeScout-1.7B remains competitive, surpassing its F1 score by **2-4%** and **3-6%** for modules and functions respectively,

Table 2: Summary of results on SWE-Bench Verified, Pro, and Lite. For all benchmarks, CodeScout achieves a new open-source SoTA, with stronger/comparable results over base and post-trained LLMs upto 8-18× larger, narrowing the gap with and often surpassing closed-source LLMs. The highest metric is bold-faced and 2^{nd} highest metric is underlined.

Benchmark	Method	File-level F1	Function-level F1
SWE-Bench Verified	RepoNavigator-7B	51.63	27.49
	CodeScout-1.7B	55.46	28.22
	RepoNavigator-32B	67.75	34.09
	CodeScout-4B	68.52	36.78
	CodeScout-14B	<u>68.57</u>	<u>40.32</u>
	RepoNavigator + GPT-5-Chat	58.88	31.17
	RepoNavigator + Claude-Sonnet-4.5	**79.94**	**43.62**
SWE-Bench Pro	RepoNavigator-7B	39.74	14.29
	CodeScout-1.7B	40.96	18.24
	RepoNavigator-32B	**57.57**	20.72
	CodeScout-4B	51.77	**29.03**
	CodeScout-14B	<u>53.63</u>	<u>28.74</u>
SWE-Bench Lite	OpenHands-Bash + Qwen3-32B (Thinking)	58.98	23.76
	CodeScout-1.7B	56.57	27.07
	CodeScout-4B	<u>67.03</u>	<u>39.87</u>
	CodeScout-14B	**71.84**	**44.43**

while trailing by **2-7%** for files. CodeScout-4B consistently outperforms **8**× larger Qwen3-32B for all benchmarks, with absolute F1 improvements of **5-8%**, **11-16%**, and **13-17%** for files, modules, and functions respectively. Finally, CodeScout-14B surpasses F1 scores of Qwen3-32B **6-13%** for files, **15-22%** for modules, and **16-21%** for functions.

5.2 CodeScout outperforms larger base and post-trained LLMs using complex agents

Comparisons with larger base LLMs using complex scaffolds: CodeScout LLMs consistently outperform Qwen2.5-32B with the OrcaLoca scaffold [39] on SWE-Bench Verified. CodeScout-4B and CodeScout-14B *significantly* exceed its F1 scores by **10%** for files and **8-11%** for functions. Impressively, our **18× smaller** CodeScout-1.7B remains competitive, trailing Qwen2.5-32B with OrcaLoca by < **0.5%** in function-level F1 and by **2.65%** in file-level F1. On SWE-bench Pro, CodeScout models exhibit more pronounced gains over Qwen2.5-32B with the CoSIL agent [14]. CodeScout-1.7B, 4B, and 14B achieve absolute F1 gains of **20-33%** for files and **11-21%** for functions.

Comparisons with larger post-trained LLMs using complex scaffolds: We primarily compare CodeScout with RepoNavigator [42] as it is our strongest post-training baseline. It is also closer to our work than RepoSearcher [20] as it is trained without relying on closed-source LLMs and can predict a variable number of locations (§4.2). For both SWE-bench Verified and Pro, CodeScout models frequently achieve superior or competitive performance over **3-8× larger** RepoNavigator models. For both benchmarks, CodeScout-1.7B achieves slightly better performance than RepoNavigator-7B with a **1-4%** higher F1 score for both files and functions. Furthermore, CodeScout-4B consistently exceeds the performance of RepoNavigator-14B across both benchmarks, with absolute F1 improvements of **2-10%** for files and **8-11%** for functions. Comparison with RepoNavigator-32B reveals distinct trends across benchmarks. On SWE-bench Verified, CodeScout-4B and CodeScout-14B achieve better function-level F1 by **3-6%** with a very similar file-level F1. On SWE-bench Pro, CodeScout-4B and CodeScout-14B outperform RepoNavigator-32B by **8%** in function-level F1, whereas the 32B baseline has **4-6%** better file-level F1 scores.

5.3 CodeScout narrows the performance gap with closed-source LLMs

On SWE-bench Verified, CodeScout-4B and CodeScout-14B LLMs outperform GPT-5-Chat with the specialized RepoNavigator agent, with F1 gains of **9%** for files and **5-9%** for functions. CodeScout-4B and 14B LLMs have impressive function-level localization performance surpassing F1 score of Claude-3.7-Sonnet with RepoNavigator by **5-8%**. Interestingly, Claude-Sonnet-4.5 achieves stronger performance with OpenHands-Bash over RepoNavigator, with **18%** higher function-level F1 and slightly better file-level F1 (2%). Similarly, GPT-5 with our agent outperforms Claude 3.7 Sonnet with RepoNavigator, with absolute F1 gains of **23%** for functions and **5%** for files. Thus, specialized scaffold design may not always improve performance. Soni et al. [30], SWE-Agent Team [31] report similar findings for other domains. Notably, GPT-5 and Claude-Sonnet-4.5 consistently outperform CodeScout when using OpenHands-Bash with the submission reminder (§4.2). However, both LLMs are surprisingly sensitive to prompt engineering. Without the reminder, their performance drops to ≈**0** on all benchmarks. Interestingly, our preliminary experiments with GPT-4o does not face this issue despite being an older proprietary LLM.

6 Analysis

This section presents detailed analysis on the benefits of effective localization on issue resolution (§6.1) and the evolving tool-use behaviors during RL (§6.2). Appendix F includes detailed error analysis of CodeScout models.

Table 3: Results on SWE-Bench Verified. CodeScout outperforms upto 8× larger open-source LLMs, competing with/surpassing proprietary LLMs. Qwen2.5 results use instruct LLMs. ^{rem}uses submission reminder (§4.2), †trained with RL; ‡RFT from CodeScout-14B; $^{\triangle}$RFT from Claude-3.7-Sonnet. Best metric is bold-faced and 2^{nd} best is underlined.

Scaffold	LLM	File-level			Module-level			Function-level		
		F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
Closed-Source LLMs										
RepoSearcher	Claude-3.7-Sonnet	32.30	20.24	**89.24**	-	-	-	26.91	18.64	**66.08**
RepoNavigator	GPT-5-Chat	58.88	61.87	58.17	-	-	-	31.17	34.56	30.42
	Claude-3.7-Sonnet	73.01	75.95	72.26	-	-	-	31.72	34.43	31.03
	Claude-Sonnet-4.5	79.94	81.92	80.68	-	-	-	43.62	45.76	43.97
OpenHands-Bash	GPT-5	3.20	3.20	3.20	2.60	2.60	2.60	2.60	2.60	2.60
	Claude-Sonnet-4.5	0.80	0.80	0.80	0.40	0.40	0.40	0.40	0.40	0.40
OpenHands-Bash^{rem}	GPT-5	78.18	79.25	80.80	61.17	62.23	63.35	54.79	56.80	56.53
	Claude-Sonnet-4.5	**82.01**	**84.50**	82.86	**67.19**	**70.11**	**67.47**	**61.78**	**65.42**	61.99
Open-Source LLMs										
CoSIL	Qwen2.5-32B	30.77	19.34	83.50	-	-	-	22.11	14.85	55.38
Agentless		35.38	25.60	78.93	-	-	-	27.33	24.07	40.97
LocAgent		44.18	34.18	79.39	-	-	-	21.48	16.29	46.79
OrcaLoca		58.11	59.51	59.57	-	-	-	28.72	25.59	39.14
RepoSearcher	Qwen2.5-7B$^{\triangle\dagger}$	30.09	18.80	83.11	-	-	-	25.57	17.68	62.38
	Qwen2.5-32B$^{\triangle\dagger}$	32.25	20.24	**88.59**	-	-	-	28.03	19.36	**68.55**
RepoNavigator	Qwen2.5-7B†	51.63	53.83	50.62	-	-	-	27.49	30.34	26.69
	Qwen2.5-14B†	58.90	58.97	61.60	-	-	-	29.23	30.08	31.02
	Qwen2.5-32B†	67.75	70.76	67.29	-	-	-	34.09	37.19	33.71
OpenHands-Bash	Qwen3-1.7B	2.40	2.09	3.60	0.37	0.32	0.60	0.34	0.32	0.50
	Qwen3-4B-Instruct	49.73	49.69	53.34	19.32	19.86	20.15	13.27	14.17	13.74
	Qwen3-14B	43.13	36.49	71.20	22.86	20.40	33.04	16.08	14.51	23.58
	Qwen3-32B (Thinking)	62.91	59.87	73.63	34.69	33.85	39.46	23.99	24.22	26.86
	CodeScout-1.7B-RFT‡	46.60	48.60	45.82	29.79	31.60	29.13	23.04	25.30	22.32
	CodeScout-1.7B‡†	55.46	58.40	54.27	36.45	39.37	35.46	28.22	31.77	27.18
	CodeScout-4B†	68.52	**71.53**	67.74	45.97	49.70	44.97	36.78	40.71	35.72
	CodeScout-14B†	**68.57**	71.00	68.69	**50.88**	**53.71**	**50.88**	**40.32**	**43.74**	40.27

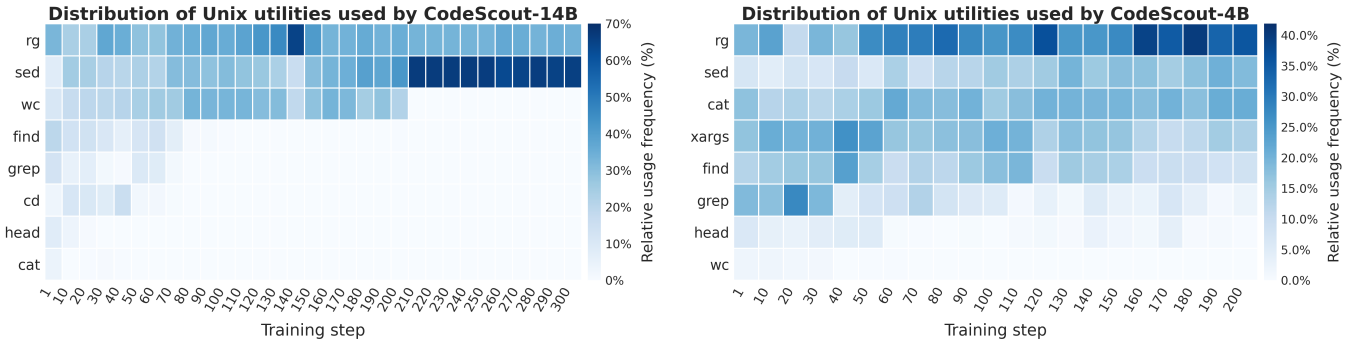
6.1 Does Effective Code Localization Improve Issue Resolution?

We show that augmenting OpenHands issue resolution agent [32] with relevant code retrieved by CodeScout improves performance on SWE-Bench Verified. We consider three settings: (1) a vanilla baseline without localization, (2) augmenting the agent with locations retrieved by CodeScout-14B, and (3) augmenting the agent with oracle locations parsed from the gold patch (§3.1). For (2) and (3), the user prompt is modified to include the names of localized files, modules, and functions (Appendix K). We evaluate Qwen3-4B-Instruct-2507 and Qwen3-Coder-30B-A3B-Instruct compare the performance of the three settings using issue resolution rate, and their efficiency using instance-level averages of the number of steps per trajectory and the total input and output tokens accumulated across all turns. Table 6 presents the detailed results.

For Qwen3-4B-Instruct, augmenting the agent with locations retrieved by CodeScout-14B increases the resolution rate by **3.8%**, while reducing trajectory length by **2.18** steps, and input and output tokens by **17.46%** and **6.71%** respectively. Similar efficiency gains are observed for Qwen3-Coder-30B-A3B-Instruct, with **2.89** fewer steps and **5.33%** fewer input tokens and **2.00%** fewer output tokens, while achieving comparable performance with a small improvement

Table 4: Results on SWE-Bench Pro. CodeScout outperforms upto 8× larger base and post-trained LLMs across various agents, narrowing the gap with closed-source LLMs. Qwen2.5 results use instruct LLMs. ^{rem} adds submission reminder (§4.2), †trained with RL; ‡RFT from CodeScout-14B. Best metric is bold-faced and 2^{nd} best metric is underlined.

Scaffold	LLM	File-level			Module-level			Function-level		
		F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
Closed-Source LLMs										
OpenHands-Bash	GPT-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Claude-Sonnet-4.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OpenHands-Bash^{rem}	GPT-5	61.18	69.10	62.06	42.20	52.86	39.63	35.86	48.91	32.65
	Claude-Sonnet-4.5	**64.75**	**75.39**	**64.49**	**48.54**	**61.75**	**45.39**	**42.26**	**58.14**	**38.06**
Open-Source LLMs										
RepoSearcher	Qwen2.5-32B	3.81	2.52	9.00	-	-	-	2.31	2.46	2.52
LocAgent		19.77	0.38	25.73	-	-	-	4.30	0.17	8.72
Agentless		20.07	13.89	43.07	-	-	-	7.98	7.31	11.08
CoSIL		20.95	14.03	48.87	-	-	-	7.67	6.00	14.03
RepoNavigator	Qwen2.5-7B†	39.74	48.13	36.36	-	-	-	14.29	21.26	12.33
	Qwen2.5-14B†	49.72	58.64	46.85	-	-	-	18.06	25.25	16.05
	Qwen2.5-32B†	**57.57**	68.69	53.49	-	-	-	20.72	29.44	18.13
OpenHands-Bash	Qwen3-1.7B	0.73	0.72	1.13	0.00	0.00	0.00	0.00	0.00	0.00
	Qwen3-4B-Instruct	36.96	44.42	35.59	11.78	17.46	10.19	8.12	12.16	7.01
	Qwen3-14B	30.08	28.48	48.22	11.87	13.82	14.21	8.20	9.97	9.92
	Qwen3-32B (Thinking)	46.85	49.65	**54.55**	21.94	27.18	22.62	12.31	17.82	11.93
	CodeScout-1.7B-RFT‡	34.54	47.74	30.22	22.43	36.53	18.59	16.07	29.04	13.02
	CodeScout-1.7B‡†	40.96	56.52	35.91	25.27	40.66	20.83	18.24	32.08	14.72
	CodeScout-4B†	51.77	**68.98**	46.16	36.97	**56.05**	31.13	**29.03**	**48.65**	23.73
	CodeScout-14B†	53.63	68.81	48.81	**37.13**	53.80	**32.02**	28.74	46.09	**23.76**



(a) Tool use for CodeScout-14B

(b) Tool use for CodeScout-4B

Figure 3: Distribution of the top-8 frequent commands used by CodeScout. LLMs start with diverse utilities and converge to a small set as training proceeds. CodeScout-14B uses only ripgrep (rg) and sed, while CodeScout-4B mainly uses rg, cat, sed, and xargs.

(+0.80%). For both LLMs, replacing CodeScout predictions with oracle locations yields further performance gains (**+6.00%** for 30B and **+2.40%** for 4B). The 30B model also benefits with improved efficiency, with **1.37** fewer steps and reductions of **10.12%** and

Table 5: Results on SWE-Bench Lite. CodeScout outperforms/competes with base LLMs upto 18× larger, narrowing the gap with proprietary LLMs. Qwen2.5 results use instruct LLMs. ^{rem}uses submission reminder (§4.2), †trained with RL; ‡RFT from CodeScout-14B; $^{\triangle}$RFT from Claude-3.7-Sonnet. Best metric is bold-faced and 2^{nd} best is underlined.

Scaffold	LLM	File-level			Module-level			Function-level		
		F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
Closed-Source LLMs										
LocAgent	Claude-3.5-Sonnet	31.39	18.83	**94.16**	29.91	18.18	**86.98**	27.53	17.08	**76.61**
OpenHands-Bash	GPT-5	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.09
	Claude-Sonnet-4.5	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36
OpenHands-Bash^{rem}	GPT-5	77.77	75.73	82.48	67.86	66.09	72.63	**61.12**	59.43	67.21
	Claude-Sonnet-4.5	**81.87**	**80.17**	85.40	**69.62**	**68.87**	72.87	61.11	**61.72**	63.72
Open-Source LLMs										
OpenHands-Bash	Qwen3-1.7B	2.16	1.96	2.92	0.36	0.36	0.36	0.00	0.00	0.00
	Qwen3-4B-Instruct	47.41	43.70	55.47	14.50	13.90	16.24	8.72	8.67	9.61
	Qwen3-14B	38.63	31.30	71.90	20.10	16.74	31.57	11.73	9.88	18.13
	Qwen3-32B (Thinking)	58.98	54.26	71.53	37.20	34.64	43.98	23.76	23.11	26.89
	CodeScout-1.7B-RFT‡	45.99	45.99	45.99	34.79	35.22	34.67	24.51	25.18	24.39
	CodeScout-1.7B‡†	56.57	56.57	56.57	41.24	41.61	41.24	27.07	28.28	26.82
	CodeScout-4B†	67.03	66.61	67.88	53.10	53.47	53.65	39.87	41.59	39.96
	CodeScout-14B†	**71.84**	**71.17**	**73.36**	**59.23**	**59.18**	**60.28**	**44.43**	**45.59**	**45.13**

Table 6: Issue resolution performance on SWE-Bench Verified: augmenting the agent with relevant locations improves performance and efficiency. Best metric is bold-faced and 2^{nd} best is underlined.

Localization Approach	Resolution Rate ↑	Avg. # Steps ↓	Avg. Input Tokens ↓	Avg. Output Tokens ↓
Qwen3-4B-Instruct				
None (Vanilla)	13.40%	16.09	344.37K	2.98K
CodeScout-14B	17.20%	**13.91**	**284.26K**	**2.78K**
Oracle	**19.60%**	16.41	327.75K	3.17K
Qwen3-Coder-30B-A3B-Instruct				
None (Vanilla)	45.20%	51.00	1596.71K	13.49K
CodeScout-14B	46.00%	48.11	1511.53K	13.22K
Oracle	**52.00%**	**46.74**	**1358.60K**	**12.48K**

5.60% in input and output tokens. Overall, these results indicate that continued improvements in code localization will further enhance both the performance and efficiency of issue resolution agents.

6.2 How does the tool-use behaviour of CodeScout evolve during RL?

To analyze the evolving tool-use behaviors RL, we plot the distribution of Unix utilities invoked by the LLM in Figure 3. We parse rollouts sampled from model checkpoints at every 10 training steps and aggregate command usage across these checkpoints to identify the top-8 frequently used commands. For each checkpoint, we compute the relative usage frequency of each utility for all rollouts logged in that training step.

We observe convergence in tool usage during RL. While the 14B LLM initially uses various utilities, after 200 training steps it uses *only* on two commands: `ripgrep` (`rg`) and `sed`. Similarly, while the 4B LLM initially uses many utilities, it converges to mainly using `ripgrep`, `sed`, `cat`, `find` and `xargs`. Thus, we can further simplify the agent to a small subset of Unix utilities without needing access to the entire terminal which is crucial in security-sensitive deployments. Appendix I presents example trajectories for 4B and 14B LLMs that illustrate how these utilities are used by the agent.

7 Conclusion

Our work presents a fully open-source RL recipe for training code search agents and releases the CodeScout family, achieving state-of-the-art localization results without complex, language-specific scaffolds used by prior work. On SWE-Bench Verified, Pro, and Lite, our models outperform larger open-source LLMs across complex setups and narrow the gap with some closed-source models. We show that augmenting agents with localized context improves both efficiency and performance, and analyze fine-grained behavioral shifts during RL via command-line tool usage, highlighting the potential for even simpler scaffolds. Appendix E discusses limitations and future research directions.

References

[1] Anthropic. 2025. *Claude Code.* https://code.claude.com/docs/en/overview Agentic AI coding assistant.

[2] Shiyi Cao, Sumanth Hegde, Dacheng Li, Tyler Griggs, Shu Liu, Eric Tang, Jiayi Pan, Xingyao Wang, Akshay Malik, Graham Neubig, Kourosh Hakhamaneshi, Richard Liaw, Philipp Moritz, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. SkyRL-v0: Train Real-World Long-Horizon Agents via Reinforcement Learning.

[3] Zhaoling Chen, Robert Tang, Gangda Deng, Fang Wu, Jialong Wu, Zhiwei Jiang, Viktor Prasanna, Arman Cohan, and Xingyao Wang. 2025. LocAgent: Graph-Guided LLM Agents for Code Localization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 8697–8727. doi:10.18653/v1/2025.acl-long.426

[4] Neil Chowdhury, James Aung, Chan Jun Shern, Oliver Jaffe, Dane Sherburn, Giulio Starace, Evan Mays, Rachel Dias, Marwan Aljubeh, Mia Glaese, Carlos E. Jimenez, John Yang, Leyton Ho, Tejal Patwardhan, Kevin Liu, and Aleksander Madry. 2024. Introducing SWE-bench Verified. https://openai.com/index/introducing-swe-bench-verified/

[5] Cursor Team. 2025. Composer: Building a Fast Frontier Model with RL. https://cursor.com/blog/composer. Accessed: 2026-01-28.

[6] Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, Karmini Sampath, Maya Krishnan, Srivatsa Kundurthy, Sean Hendryx, Zifan Wang, Vijay Bharadwaj, Jeff Holm, Raja Aluri, Chen Bo Calvin Zhang, Noah Jacobson, Bing Liu, and Brad Kenstler. 2025. SWE-Bench Pro: Can AI Agents Solve Long-Horizon Software Engineering Tasks? arXiv:2509.16941 [cs.SE] https://arxiv.org/abs/2509.16941

[7] Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. 2026. AReaL: A Large-Scale Asynchronous Reinforcement Learning System for Language Reasoning. arXiv:2505.24298 [cs.LG] https://arxiv.org/abs/2505.24298

[8] Andrew Gallant. 2016. ripgrep. https://github.com/BurntSushi/ripgrep. Accessed: 2026-03-11.

[9] Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. 2025. Soft Adaptive Policy Optimization. arXiv:2511.20347 [cs.LG] https://arxiv.org/abs/2511.20347

[10] Tyler Griggs, Sumanth Hegde, Eric Tang, Shu Liu, Shiyi Cao, Dacheng Li, Charlie Ruan, Philipp Moritz, Kourosh Hakhamaneshi, Richard Liaw, Akshay Malik, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Evolving SkyRL into a Highly-Modular RL Framework. Notion Blog.

[11] Kelly Hong, Anton Troynikov, and Jeff Huber. 2025. *Context Rot: How Increasing Input Tokens Impacts LLM Performance.* Technical Report. Chroma. https://research.trychroma.com/context-rot

[12] Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. 2019. Codesearchnet challenge: Evaluating the state of semantic code search. *arXiv preprint arXiv:1909.09436* (2019).

[13] Naman Jain, Jaskirat Singh, Manish Shetty, Liang Zheng, Koushik Sen, and Ion Stoica. 2025. R2E-Gym: Procedural Environments and Hybrid Verifiers for Scaling Open-Weights SWE Agents. arXiv:2504.07164 [cs.SE] https://arxiv.org/abs/2504.07164

[14] Zhonghao Jiang, Xiaoxue Ren, Meng Yan, Wei Jiang, Yong Li, and Zhongxin Liu. 2025. Issue Localization via LLM-Driven Iterative Code Graph Searching. arXiv:2503.22424 [cs.SE] https://arxiv.org/abs/2503.22424

[15] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? arXiv:2310.06770 [cs.CL] https://arxiv.org/abs/2310.06770

[16] Myeongsoo Kim, Shweta Garg, Baishakhi Ray, Varun Kumar, and Anoop Deoras. 2026. CodeAssistBench (CAB): Dataset Benchmarking for Multi-turn Chat-Based Code Assistance. arXiv:2507.10646 [cs.SE] https://arxiv.org/abs/2507.10646

[17] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding R1-Zero-Like Training: A Critical Perspective. arXiv:2503.20783 [cs.LG] https://arxiv.org/abs/2503.20783

[18] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG] https://arxiv.org/abs/1711.05101

[19] Michael Luo, Naman Jain, Jaskirat Singh, Sijun Tan, Ameen Patel, Qingyang Wu, Alpay Ariyak, Colin Cai, Tarun Venkat, Shang Zhu, Ben Athiwaratkun, Manan Roongta, Ce Zhang, Erran Li Li, Raluca Ada Popa, Koushik Sen, and Ion Stoica. 2025. DeepSWE: Training a Fully Open-sourced, State-of-the-Art Coding Agent by Scaling RL. https://www.together.ai/blog/deepswe. Together AI Blog.

[20] Zexiong Ma, Chao Peng, Qunhong Zeng, Pengfei Gao, Yanzhen Zou, and Bing Xie. 2025. Tool-integrated Reinforcement Learning for Repo Deep Search. arXiv:2508.03012 [cs.SE] https://arxiv.org/abs/2508.03012

[21] Siru Ouyang, Wenhao Yu, Kaixin Ma, Zilin Xiao, Zhihan Zhang, Mengzhao Jia, Jiawei Han, Hongming Zhang, and Dong Yu. 2025. RepoGraph: Enhancing AI Software Engineering with Repository-level Code Graph. arXiv:2410.14684 [cs.SE] https://arxiv.org/abs/2410.14684

[22] Ben Pan, Carlo Baronio, Pietro Marsella Albert Tam, Mokshit Jain, Daniel Chiu, Swyx, and Silas Alberti. 2025. Introducing SWE-grep and SWE-grep-mini: RL for Multi-Turn, Fast Context Retrieval. https://cognition.ai/blog/swe-grep. Blog post.

[23] Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. 2025. Training Software Engineering Agents and Verifiers with SWE-Gym. arXiv:2412.21139 [cs.SE] https://arxiv.org/abs/2412.21139

[24] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2026. YaRN: Efficient Context Window Extension of Large Language Models. arXiv:2309.00071 [cs.CL] https://arxiv.org/abs/2309.00071

[25] Weihan Peng, Yuling Shi, Yuhang Wang, Xinyun Zhang, Beijun Shen, and Xiaodong Gu. 2025. SWE-QA: Can Language Models Answer Repository-level Code Questions? arXiv:2509.14635 [cs.CL] https://arxiv.org/abs/2509.14635

[26] SID Research. 2025. SID-1 Technical Report: Test-Time Compute for Retrieval. *SID AI* (2025). https://www.sid.ai/research/SID-1-technical-report.

[27] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3 (2009), 333–389. https://api.semanticscholar.org/CorpusID:207178704

[28] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300 [cs.CL] https://arxiv.org/abs/2402.03300

[29] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. HybridFlow: A Flexible and Efficient RLHF Framework. In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys '25).* ACM, 1279–1297. doi:10.1145/3689031.3696075

[30] Aditya Bharat Soni, Boxuan Li, Xingyao Wang, Valerie Chen, and Graham Neubig. 2025. Coding Agents with Multimodal Browsing are Generalist Problem Solvers. arXiv:2506.03011 [cs.CL] https://arxiv.org/abs/2506.03011

[31] SWE-Agent Team. 2024. mini-SWE-agent. https://github.com/SWE-agent/mini-swe-agent

[32] Xingyao Wang, Simon Rosenberg, Juan Michelini, Calvin Smith, Hoang Tran, Engel Nyst, Rohit Malhotra, Xuhui Zhou, Valerie Chen, Robert Brennan, and Graham Neubig. 2025. The OpenHands Software Agent SDK: A Composable and Extensible Foundation for Production Agents. arXiv:2511.03690 [cs.SE] https://arxiv.org/abs/2511.03690

[33] Zora Zhiruo Wang, Akari Asai, Frank F Xu, Yiqing Xie, Graham Neubig, Daniel Fried, et al. 2025. Coderag-bench: Can retrieval augment code generation?. In *Findings of the Association for Computational Linguistics: NAACL 2025.* 3199–3214.

[34] Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. 2024. Agentless: Demystifying llm-based software engineering agents. *arXiv preprint arXiv:2407.01489* (2024).

[35] 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. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 1123–1139. doi:10.18653/v1/2025.findings-acl.62

[36] 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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[37] John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. arXiv:2405.15793 [cs.SE] https://arxiv.org/abs/2405.15793

[38] John Yang, Kilian Lieret, Carlos E. Jimenez, Alexander Wettig, Kabir Khandpur, Yanzhe Zhang, Binyuan Hui, Ofir Press, Ludwig Schmidt, and Diyi Yang. 2025. SWE-smith: Scaling Data for Software Engineering Agents. arXiv:2504.21798 [cs.SE] https://arxiv.org/abs/2504.21798

[39] Zhongming Yu, Hejia Zhang, Yujie Zhao, Hanxian Huang, Matrix Yao, Ke Ding, and Jishen Zhao. 2025. OrcaLoca: An LLM Agent Framework for Software Issue Localization. arXiv:2502.00350 [cs.SE] https://arxiv.org/abs/2502.00350

[40] 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. arXiv:2308.01825 [cs.CL] https://arxiv.org/abs/2308.01825

[41] 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.

arXiv:2504.02605 [cs.SE] https://arxiv.org/abs/2504.02605
[42] Zhaoxi Zhang, Yitong Duan, Yanzhi Zhang, Yiming Xu, Zhixiang Wang, Kun Liang, Yang Li, Jiahui Liang, Deguo Xia, Jizhou Huang, et al. 2025. One Tool Is Enough: Reinforcement Learning for Repository-Level LLM Agents. *arXiv preprint arXiv:2512.20957* (2025).
[43] Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. 2023. Click: Controllable Text Generation with Sequence Likelihood Contrastive Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1022–1040. doi:10.18653/v1/2023.findings-acl.65
[44] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group Sequence Policy Optimization. arXiv:2507.18071 [cs.LG] https://arxiv.org/abs/2507.18071

A Prompts and Tool Definitions for OpenHands-Bash

Figures 4, 5 describe the system prompt, and Figure 6 includes the user prompt, used for all LLMs (including CodeScout) trained or evaluated with OpenHands-Bash (§3.2). The system prompt specifies the `max_turns` to be 6 for CodeScout-4B and its base model, and to 4 for all other LLMs. Table 7 provides the schemas for the tools used by OpenHands-Bash.

B Detailed experimental settings for training and evaluation

B.1 Experimental setup for training runs

This section details the experimental setup for all our training runs from §4.1 used for training CodeScout models. Note that when processing the ground truth patches in our training data we enhance the patch processing scripts from LocAgent [3] to (i) detect additions of member functions and class attributes at the module and file level, (ii) capture modifications to import statements and global variables at the file level, and (iii) ignore edits to docstrings within functions and classes.

We train CodeScout-4B for 200 steps on 1.6K instances using a batch size of 8 and sample 8 rollouts per instance. The model uses a maximum context length of 40K tokens, the reward function from Equation 1, and a maximum of 6 turns per episode. We train CodeScout-14B for 300 steps on 9.6K instances with a batch size of 32, sampling 4 rollouts per instance. We use a maximum context length of 50K tokens, extended with YaRN [24], and the reward function from Equation 1 augmented with the auxiliary term $r^{\text{turn}}(\tau, k)$ while k is set to 4 as the agent is allowed a maximum of 4 turns per rollout.

To warm-start the 1.7B model with rejection sampling fine-tuning (RFT) [40] on trajectories sampled from CodeScout-14B, we sample rollouts from CodeScout-14B on a random subset of 7.7K training instances and retain only those achieving perfect localization across all three granularities (i.e., F1 = 1.0 for file, module, and function), yielding 4K training examples. We then use the veRL framework [29] to perform supervised fine-tuning of `Qwen3-1.7B` on these successful trajectories for one epoch, with a learning rate of $5e^{-5}$, a cosine learning rate scheduler, warmup ratio of 0.1, global batch size of 8, and the AdamW optimizer [18]. The resulting checkpoint (CodeScout-1.7B-RFT) is subsequently trained using RL (§3.4) for 100 steps on 800 instances (not seen during RFT) with a batch size of 8, sampling 8 rollouts per instance. We use 32K context window, the reward from Equation 1, and allow upto 4 turns per episode.

Across all CodeScout RL training runs, we use a constant learning rate of $1e^{-6}$, set `clip_ratio_low` to $3e^{-4}$ and `clip_ratio_high` to $4e^{-4}$, use the AdamW optimizer, and sample rollouts with temperature of 1.0. Note that we use a modified chat template for 1.7B and 14B models that preserves the `<think>` and `</think>` tokens from earlier turns during tokenization (https://huggingface.co/OpenPipe/Qwen3-14B-Instruct), overriding the default behavior that removes them. This modification preserves the sequence extension property, where previous trajectory steps are guaranteed to be prefixes for future steps allowing us to merge all steps into a single training sequence, greatly improving efficiency. All experiments are conducted with 8×H100 GPUs. Appendix C presents training curves for RL rewards and SFT loss.

B.2 Experimental setup for evaluation

This section describes the experimental setup for all the evaluation runs performed in §4.2 and §6.1. During evaluation of localization performance of Qwen3 models, as well as during rollout generation from CodeScout-14B for training CodeScout-1.7B-RFT (§4.1), we use the decoding hyperparameters recommended by the Qwen3 developers. For instruct and reasoning models with thinking disabled, we use temperature = 0.7, top-k = 20, and top-p = 0.8. For reasoning models with thinking enabled, we use temperature = 0.6, top-k = 20, and top-p = 0.95. Finally, we set the maximum context length to 132K tokens for all Qwen3 models. For experiments in §6.1, our models are evaluated using a 128K token context window, allowing upto 100 turns, and using the recommended decoding hyperparameters from Qwen3 developers.

C Reward and Loss Curves for CodeScout

This section presents the RL reward curves and SFT loss curves from the training runs of CodeScout. We present the aggregate reward (computed as the sum of the F1 scores across the three localization granularities for 4B and 1.7B, and for 14B as the sum of these three F1 scores and the auxiliary binary reward) and the individual file-level, module-level, and function-level F1 scores during training. Note that all reward curves are smoothed using a running average with a window of 16 training steps. Figures 7, 8, and 9 present these reward curves for CodeScout-14B, CodeScout-4B, and CodeScout-1.7B respectively. Furthermore, Figure 10 presents the loss curve for rejection sampling fine-tuning of `Qwen3-1.7B` resulting in CodeScout-1.7B-RFT model.

As expected, all reward curves across all models show an approximately increasing nature (or remain constant during the later stages of training). Notably, all the reward curves for CodeScout-4B and the CodeScout-14B models show an increasing trend with no visible signs of saturation, indicating that training further on more GitHub issues will likely result in even stronger localization performance. However, we observe that the reward curves for CodeScout-1.7B have mostly plateaued and the model no longer shows signs of further improvement with more training. Interstingly, even after fine-tuning `Qwen3-1.7B` on **4K successful trajectories** sampled from CodeScout-14B with the training loss approaching *almost zero* (Figure 10), we find that training this checkpoint further

System Prompt for the OpenHands-Bash agent used by CodeScout (continued on next page)

```
You are a specialized code localization agent. Your sole objective is to identify and return the files in the codebase
      that are relevant to the user's query.
You are given access to the codebase in a linux file system.

## PRIMARY DIRECTIVE
- Find relevant files, do NOT answer the user's query directly
- Prioritize precision: every file you return should be relevant
- You have up to {{ max_turns }} turns to explore and return your answer

## TOOL USAGE REQUIREMENTS

### bash tool (REQUIRED for search)
- You MUST use the bash tool to search and explore the codebase
- Execute bash commands like: rg, grep, find, ls, cat, head, tail, sed
- Use parallel tool calls: invoke bash tool up to 5 times concurrently in a single turn
- NEVER exceed 5 parallel tool calls per turn
- Common patterns:
  * `rg "pattern" -t py` - search for code patterns
  * `rg --files | grep "keyword"` - find files by name
  * `cat path/to/file.py` - read file contents
  * `find . -name "*.py" -type f` - locate files by extension
  * `wc -l path/to/file.py` - count lines in a file
  * `sed -n '1,100p' path/to/file.py` - read lines 1-100 of a file
  * `head -n 100 path/to/file.py` - read first 100 lines
  * `tail -n 100 path/to/file.py` - read last 100 lines

### Reading Files (CRITICAL for context management)
- NEVER read entire large files with `cat` - this will blow up your context window
- ALWAYS check file size first: `wc -l path/to/file.py`
- For files > 100 lines, read in chunks:
  * Use `sed -n '1,100p' file.py` to read lines 1-100
  * Use `sed -n '101,200p' file.py` to read lines 101-200
  * Continue with subsequent ranges as needed (201-300, 301-400, etc.)
- Strategic reading approach:
  * Read the first 50-100 lines to see imports and initial structure
  * Use `rg` to find specific patterns and their line numbers
  * Read targeted line ranges around matches using `sed -n 'START,ENDp'`
  * Only read additional chunks if the initial sections are relevant

### Submitting Your Answer (REQUIRED)

When you have identified all relevant locations, you MUST use the `localization_finish` tool to submit your results.

**When to include what:**
1. If the required modifications belong to a specific function that belongs to a class, provide the file path, class
     name, and function name.
2. If the required modification belongs to a function that is not part of any class, provide the file path and
     function name.
3. If the required modification does not belong to any specific class or a function (e.g. global variables, imports,
     new class, new global function etc.), it is sufficient to provide only the file path.
```

Figure 4: System prompt for the OpenHands-Bash agent scaffold used by CodeScout (continued on next page).

System Prompt for the OpenHands-Bash agent used by CodeScout

```
4. If the required modification belongs to a class (e.g. adding a new method to a class), provide the file path and
     class name.

## SEARCH STRATEGY
1. **Initial Exploration**: Cast a wide net
   - Search for keywords, function names, class names
   - Check file names and directory structure
   - Use up to 3 parallel bash calls to explore multiple angles
   - Check file sizes with `wc -l` before reading
   - Read promising files in chunks (lines 1-100) to verify relevance

2. **Deep Dive**: Follow the most promising leads
   - Use up to 3 parallel bash calls to investigate further
   - Read files in chunks to confirm they address the query
   - Use `rg` with line numbers to locate specific code, then read those ranges
   - Start eliminating false positives

3. **Final Verification**: Confirm your location list and terminate execution by calling the `localization_finish`
     tool
## CRITICAL RULES
- NEVER exceed 5 parallel bash tool calls in a single turn
- ALWAYS use the `localization_finish` tool after identifying all relevant locations
- ALWAYS use bash tool to search (do not guess file locations)
- NEVER read entire large files - always read in chunks (100-line ranges)
- Check file size with `wc -l` before reading
- Read file contents in chunks to verify relevance before including them
- Return file paths as they appear in the repository. Do not begin the path with "./"
- Aim for high precision (all files relevant) and high recall (no relevant files missed)
- Class and function names are OPTIONAL - only include when changes are at that level
## EXAMPLE OUTPUT BEHAVIOUR
Here are some examples of how to format your output when calling the `localization_finish` tool:
- src/parsers/parser.py requires changes to imports, a function parse_data which belongs to the class DataParser, and
     another function __str__ inside the same class. This should be represented as three separate entries: one with
     just the file path and one each for the two functions parse_data and __str__ with file path, class name, and
     function name.
- src/user.py requires changes to a global function get_user outside of any class. This should be represented as a
     single entry with file path and function name.
- utils/visualizer.py requires adding new function visualize inside the class Visualizer. This should be represented
     as a single entry with file path and class name.
- utils/configs/default_config.py requires adding a new global function and a new class. This should be represented as
      a single entry with just the file path. Do NOT include class or function names for this file since multiple
     implementations might be possible with different function names.
```

Figure 5: System prompt for the OpenHands-Bash scaffold used by CodeScout.

using reinforcement learning improves the model performance even further - we observe an increasing reward curve with a steep slope during the first ≈ 20 training steps before saturating.

D Detailed Overview of Related Work

D.1 Approaches for code localization

Several prior methods have developed methods targeting code localization for downstream issue resolution in repositories. Jimenez et al. [15] identify relevant source code files using BM25 retrieval [27] treating the issue description as the query and the Python source

User prompt for the OpenHands-Bash agent used by CodeScout

```
I have access to a python code repository in the directory {{ working_dir }} . Consider the following issue
    description:

<issue_description>
{{ instance.problem_statement }}
</issue_description>

Act as a code search agent and localize the specific files, classes or functions of code that need modification to
    resolve the issue in <issue_description>.

NOTE: You do not need to solve the issue, all you need to do is localize relevant code from the repository. Your
    output will be used to guide another agent to solve the issue.

IMPORTANT: Your output MUST follow the below rules:
1. The final output must be a tool call to the "localization_finish" tool containing relevant code locations.
2. The locations of the file path must be RELATIVE to the {{ working_dir }} directory WITHOUT any leading "./" in the
    output.
3. Only include those locations in your output that need modification to resolve the issue in <issue_description>. Do
    NOT include any locations that do not need modification.
```

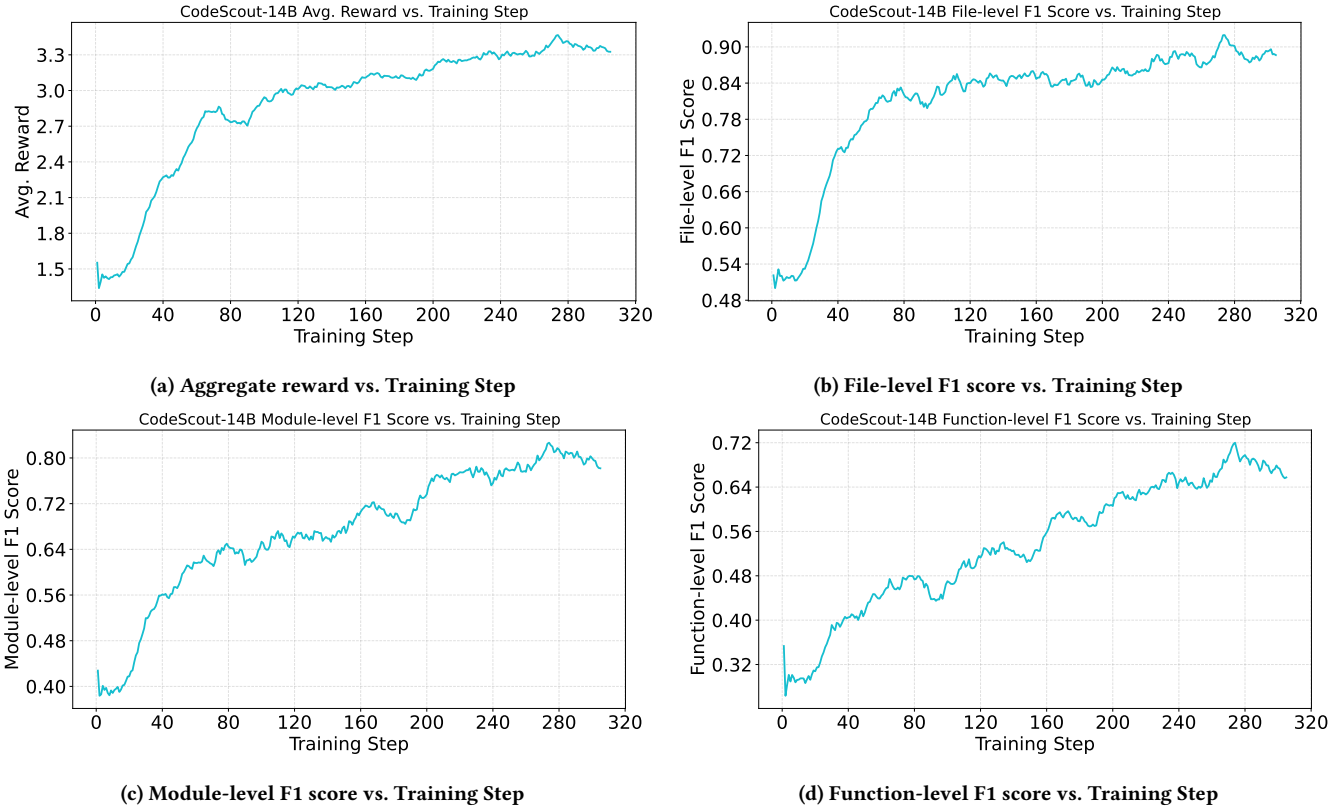
Figure 6: User prompt for the OpenHands-Bash scaffold used by CodeScout.

Table 7: Pythonic Tool Schema for the tools used by OpenHands-Bash

Tool Name	Parameters	Python Type / Constraints	Required
terminal	command	`str`	✓
	security_risk	`Literal["UNKNOWN", "LOW", "MEDIUM", "HIGH"]`	✓
	is_input	`bool`	✗
	timeout	`float`	✗
	reset	`bool`	✗
localization_finish	locations	`List[Dict[str, Optional[str]]]`	✓
	↪ file	`str`	✓
	↪ class_name	`str | None`	✗
	↪ function_name	`str | None`	✗

code files as documents. SWE-Fixer [35] use a two-step coarse-to-fine localization approach where they first retrieve relevant source code files using BM25 retrieval and then prompt a fine-tuned 7B model with the skeleton of the retrieved files to predict the names of relevant files for issue resolution. On the other hand, Agentless [34] uses a complex, multi-step localization pipeline wherein it first retrieves suspicious code files from the repository using both embedding-based retrieval and by prompting an LLM with the high-level repository structure, followed by prompting an LLM to predict relevant functions and classes given the skeletons of the suspicious files.

Many recent approaches have increasingly shifted towards developing models and systems for code localization using agentic scaffolds. Table 8 provides an overview of tools used by several prior agentic approaches and CodeScout for code localization. LocAgent [3] and RepoGraph [21] utilize a graph-based indexing approach wherein all dependencies in the code repository (for e.g. import, invoke, inherit, etc.) are captured in a code graph and also develop specialized tools allowing the agent to search and traverse the graph. Furthermore, Chen et al. [3] train a 7B model for their scaffold using rejection sampling fine-tuning on successful trajectories sampled from Claude-3.5-Sonnet. On the other hand, CoSIL [14] first localizes relevant files using module-call graphs and then identifies relevant functions using function-call graphs. Note that these call graphs are constructed on-the-fly during inference as opposed to pre-indexing the code graph done by Chen et al. [3]. OrcaLoca [39] performs code localization through efficient exploration of the code graph using priority-based action scheduling, action decomposition with relevance scoring, and distance-aware context pruning. Note that many of these scaffolds are not suitable for reinforcement learning as they often require expensive repository pre-processing (for e.g. creating a code graph) increasing the computational overhead of performing rollouts during reinforcement learning.



(a) Aggregate reward vs. Training Step

(b) File-level F1 score vs. Training Step

(c) Module-level F1 score vs. Training Step

(d) Function-level F1 score vs. Training Step

Figure 7: Reward curves for RL training of CodeScout-14B.

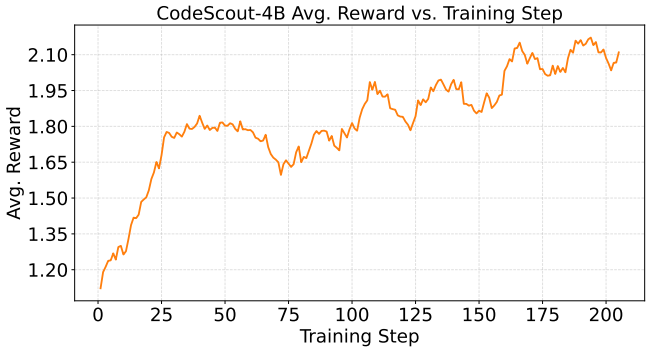
RepoSearcher [20] develop a light-weight scaffold with tools specialized for localization that allow searching for classes and functions in files, obtaining imports of a given file, etc. They use a two-stage approach to train open-source LLMs for their scaffold: (1) rejection sampling fine-tuning from a closed-source LLM (Claude3.7-Sonnet) to warm-up their model, (2) reinforcement learning on the fine-tuned checkpoint to further enhance performance. RepoNavigator [42] use an even simpler scaffold comprising with just one tool - `jump` - which allows the agent to retrieve definitions of Python symbols in files. Although Zhang et al. [42] do not rely on rejection sampling fine-tuning from closed-source LLMs and directly train models using reinforcement learning, they still rely on selecting "easy" training instances by discarding all instances that were not successfully solved by the base Qwen2.5-7B model (equipped with their scaffold) atleast once among 16 sampled trajectories.

D.2 Training methods for software engineering agents

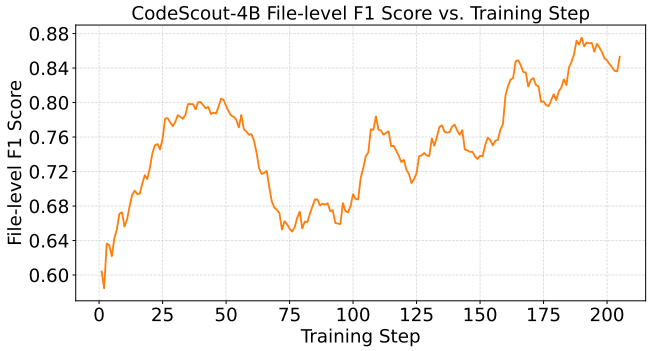
Several prior approaches have trained LLM-based software engineering agents for various downstream tasks like code localization and issue resolution. Many methods rely on performing rejection sampling fine-tuning [40]: sample trajectories (for training instances) from a stronger (often closed-source) model and train a smaller model on trajectories which successfully solve the task. Prior approaches that use this approach include LocAgent [3] and RepoSearcher [20] for code localization, and SWE-Gym [23], SWE-Smith [38], and R2E-Gym [13] for issue resolution. Recently, various attempts have been made to train models directly with reinforcement learning without relying on proprietary, closed-source LLMs for rejection sampling fine-tuning. While RepoNavigator [42], SWE-Grep [22], and SID-1 [26] train LLM agents with RL for code localization, DeepSWE [19] and SkyRL-v0 [2] leverage RL to train LLM agents for issue resolution.

E Limitations and Directions for Future Work

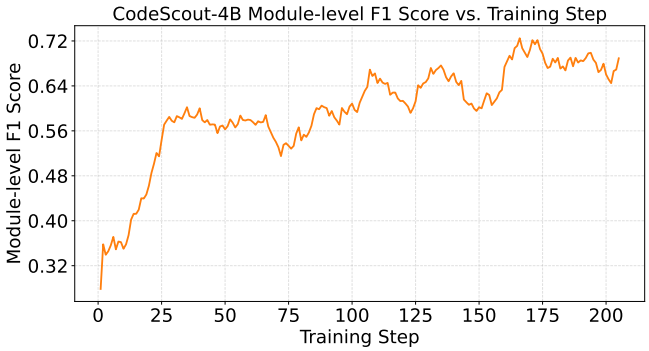
Although the proposed CodeScout recipe is significantly more general and scalable than most prior approaches for code localization, several important limitations remain. First, despite our scaffold being inherently programming language-agnostic by design, making it a potentially better candidate for cross-language generalization, our training and evaluation experiments are largely restricted to Python repositories. This constraint arises primarily from the lack of large-scale training counterparts to evaluation benchmarks such as Multi-SWE-Bench [41], containing issues from repositories written in languages other than Python. In addition, extracting ground-truth locations still requires programming language-specific processing of ground truth patches, which not only limits our ability to train agents for other programming languages, but also evaluate zero-shot transfer to other languages. Thus, we leave it for



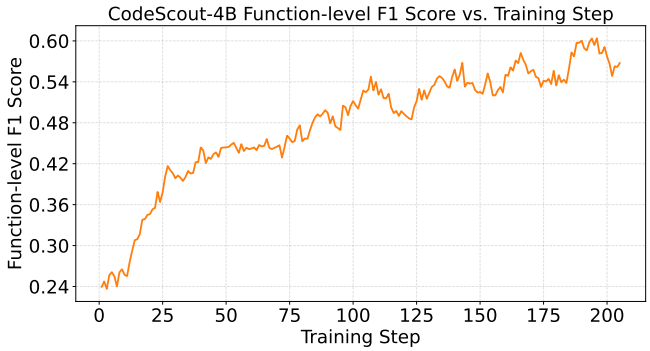
(a) Aggregate reward vs. Training Step



(b) File-level F1 score vs. Training Step



(c) Module-level F1 score vs. Training Step



(d) Function-level F1 score vs. Training Step

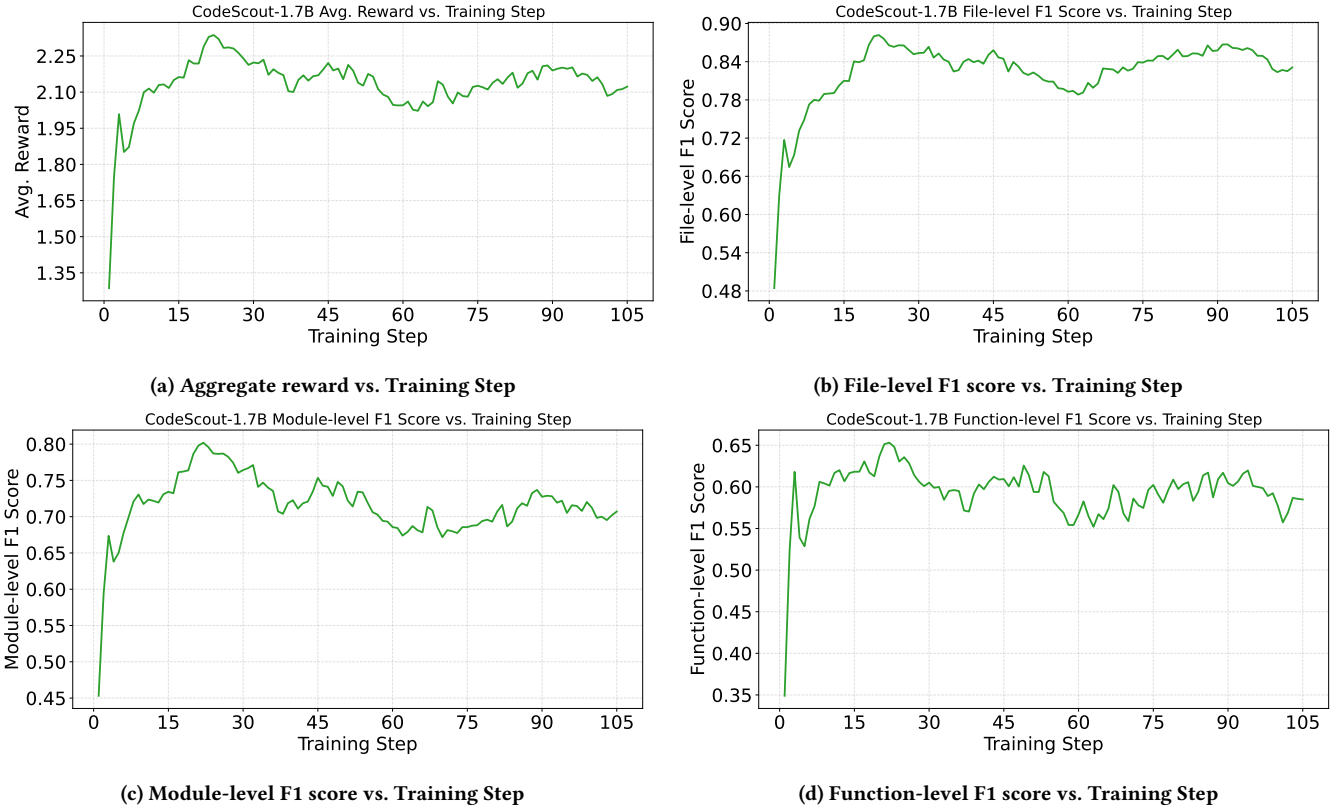
Figure 8: Reward curves for RL training of CodeScout-4B.

future work to curate training datasets and develop enhanced patch processing methods for extracting ground truth locations that will cover a broader range of programming languages.

Another limitation is that our ground truth localization target is primarily derived from the modified regions of the ground truth issue resolution code patch. However, note that there may be multiple valid patches that can resolve an issue that our recipe does not consider and effectively penalizes the agent for predicting locations from these alternative issue resolution patches. Furthermore, our data curation methodology omits other code segments that are relevant for resolving an issue but do not require modification. Note that all these limitations also exist in all the baselines we compare CodeScout against. Moreover, it is non-trivial to address them because we need to curate data at a training scale which makes this problem even harder to fix. First, all the issues in our training and evaluation datasets provide a single ground truth issue resolution patch without any alternative solutions. Second, determining relevance of a code location (that has not been modified) for resolving a given GitHub issue is somewhat subjective and difficult to determine without in-depth understanding of the repository. Thus, we leave it to future work to develop better data curation methodologies to address these limitations.

Thirdly, we acknowledge that several baselines use models from the Qwen2.5 series, while CodeScout uses the Qwen3 series, thereby our results compare scaffolds and RL recipes with different LLM backbones. However, note that our work makes significant efforts to compare CodeScout against numerous competitive baselines to ensure fair comparisons. Firstly, we include comparisons of CodeScout with significantly larger Qwen3 base models when using the same scaffold (§5.1). Secondly, when comparing with prior complex scaffolds that use models from Qwen2.5 series, we primarily compare CodeScout against **2-18**× larger LLMs (§5.2). Finally, we also include comparisons with numerous frontier closed-source LLMs, both when using specialized agents and our OpenHands-Bash scaffold (§5.3). We argue that our experimental setup is already very comprehensive and we leave it to future work to run additional RL experiments with these specialized agent frameworks equipped with the latest model series, given the high cost of running RL experiments.

Furthermore, while our reward design (§3.3) is quite intuitive and incorporates localization performance at three distinct granularities (files, modules, and functions), and a turn-limit reward for the 14B LLM, we do not study more sophisticated approaches to aggregate these individual components instead of simply summing them with equal weights. Alternative reward designs may allocate different weights to these individual components, which may result in better localization performance. However, we do not perform this analysis because determining these individual weights is non-trivial and often requires numerous ablation experiments for RL training that are very expensive to perform.



(a) Aggregate reward vs. Training Step

(b) File-level F1 score vs. Training Step

(c) Module-level F1 score vs. Training Step

(d) Function-level F1 score vs. Training Step

Figure 9: Reward curves for RL training of CodeScout-1.7B.

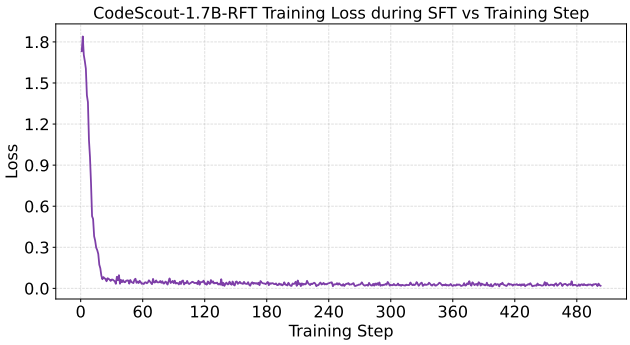


Figure 10: Loss curve for rejection sampling finetuning for CodeScout-1.7B-RFT.

Finally, localizing relevant/faulty code from the repository for resolving GitHub issues does not cover the broader scope of repository-level code search tasks, and our work does not address general-purpose repository-level question-answering/search tasks like those covered by benchmarks like CodeAssistBench [16] and SWE-QA [25]. Future research directions include expanding the scope of our work to a broader range of tasks requiring repository-level code localization.

F Error Analysis

This sections presents insights and key takeaways from our quantitative and qualitative error analysis of agent trajectories for evaluation benchmarks.

F.1 How does the localization performance vary with number of ground truth locations edited in the issue resolution patch?

We analyze the effect of number of locations (files and functions) edited in the ground truth issue resolution patch on localization performance of different LLMs when using the OpenHands-Bash scaffold. This analysis is performed with the 14B, 4B and 1.7B variants of CodeScout, and we also include the top-3 strongest performing LLMs - Claude-Sonnet-4.5, GPT-5, and Qwen3-32B (with thinking enabled). We consider the Python-only subset of SWE-Bench Pro [6] since it is significantly more challenging than SWE-Bench Verified [4] and SWE-Bench Lite [15]. Concretely, for each model, we group instances based on the number of locations (files or functions) in the ground truth localization targets, and compute the average F1 score of the agent on instances in each of these groups, and plot the trends for each of these granularities (file-level and function-level) in Figure 11.

For all LLMs, we observed that the localization performance generally decreases as the number of locations edited by the ground

Table 8: Comparison of tools used by prior code localization agents and CodeScout. Existing approaches rely on specialized task-specific tools, while CodeScout achieves competitive or superior performance using only a bash terminal typical of coding agents.

Method	Tools	Tool Behaviour
LocAgent [3]	`SearchEntity`	Performs keyword-based retrieval to identify relevant code entities.
	`TraverseGraph`	Traverse/parse codebase dependencies by navigating hops in the code graph.
	`RetrieveEntity`	Accesses the complete implementation for a specific entity ID.
CoSIL [14]	`search_class_node`	Retrieves the full code snippet of a class given its file path.
	`search_class_function_node`	Extracts implementation details for a specific member function in a class.
	`search_file_function_node`	Fetches the source code for standalone/static functions within a file.
OrcaLoca [39]	`search_file_contents`	Returns file contents or a structural skeleton for long files (>200 lines).
	`search_class`	Locates and returns class definitions or their structural skeletons.
	`search_method_in_class`	Extracts specific method implementations from within a designated class.
	`search_callable`	Identifies and returns code snippets for callable objects (functions/methods).
	`search_source_code`	Performs a general search to match source code strings to snippets.
RepoSearcher [20]	`GetRepoStructure`	Provides an overview of the repository's file and directory hierarchy.
	`GetImportOfFile`	Identifies and lists all imports for a given file.
	`SearchClass`	Fetches the raw code content of a specific class definition.
	`SearchFunction`	Fetches the raw code content of a standalone function definition.
	`SearchClassMethod`	Specifically targets and retrieves implementations of class-specific methods.
RepoNavigator [42]	`jump`	Navigates directly to a symbol's definition using a language server.
CodeScout *(Ours)*	`terminal`	Executes standard Unix commands in a persistent, stateful `tmux` session

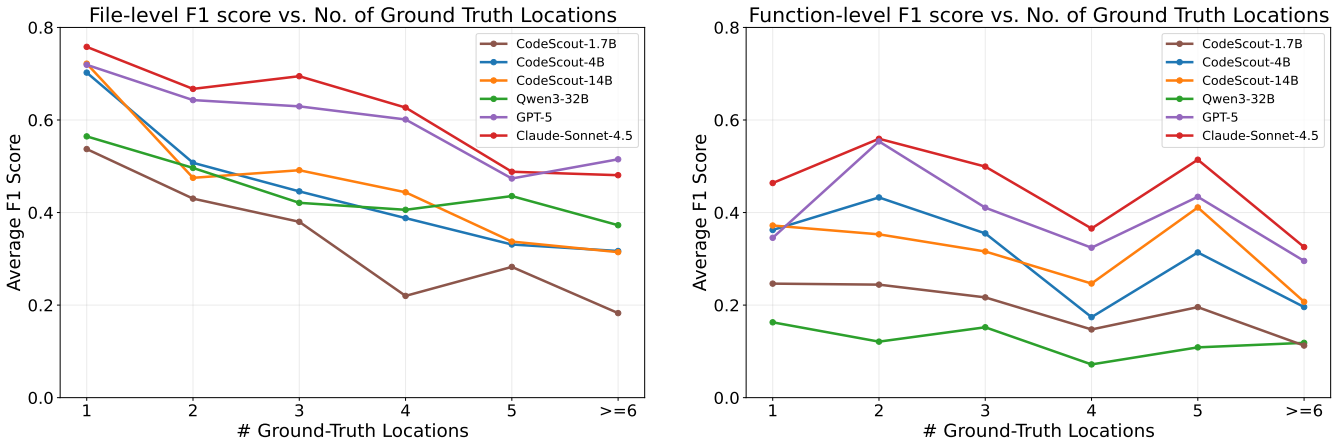


Figure 11: Average F1 score versus the number of locations edited by the ground truth issue resolution patch at the file-level (left) and function-level (right) on the SWE-Bench Pro dataset. The performance of all models drops as issues grow more complex with more files/functions being edited by the issue resolution patch.

truth issue resolution patch increases, and this trend is especially evident for file-level localization. Interestingly, for file-level localization, we observe that the performance gap between the frontier, closed-source LLMs and CodeScout-14B and 4B models widens with increasing number of ground truth files, implying that our models still lag behind proprietary LLMs when localizing context for issues that require editing more files. A similar trend is observed when comparing Qwen3-32B and CodeScout-1.7B for file-level localization. Finally, we observe that rate of drop in model performance is much higher for the CodeScout models as compared to Qwen3-32B. We hypothesize that this because our training dataset

(SWE-Smith) does not include many complex issues that edit several files as a result of which the model may not generalize too well on such issues.

Moreover, we observe that the drop in performance with increasing number of functions edited in the ground truth has a much smaller slope as compared to that for file-level localization for all LLMs. A potential reason for this observation is that many issues require editing multiple functions in the smaller set of file(s) and LLMs can identify these relatively more easily once the right set of files are localized. Similar to file-level localization, we observe that closed-source LLMs generally achieve much better localization performance over CodeScout-14B and 4B models for issues that edit a larger number of functions compared to the performance gap on issues that edit fewer functions. However, when comparing CodeScout-1.7B with Qwen3-32B, we find that the 1.7B consistently achieves stronger or competitive performance over **18x** larger base LLM regardless of the number of ground truth functions. This trend contrasts against that observed for file-level localization for these models implying that LLMs from the Qwen3 family greatly benefit from our RL recipe for function-level localization and base LLMs perform much poorly at this task.

F.2 Qualitative error analysis

To understand the qualitative error behaviours of CodeScout models when performing code localization, we analyze agent trajectories on the subset of examples from SWE-Bench Pro. In particular, we choose 20 successful trajectories (where the total F1 score for the three granularities is >= 2.4) and 20 failure trajectories (where the total F1 score for the three granularities is <= 0.3), and identify model behaviours responsible for success/failure during localization. We perform this analysis for CodeScout-1.7B, 4B, and 14B models. Across all three model scales, our analysis reveals consistent behavioral contrasts between successful and failed localization attempts as discussed below. Note that the general problem-solving strategy for all these models is to first perform a broader keyword-based search to identify the right set of files to view, and then read specific line ranges from these files to determine the relevant set of classes and functions.

Search specificity: We find that the ability of the model to search for precise keywords in the beginning of the trajectory is the strongest differentiator - successful trajectories are characterized by precise search terms like specific function or class names (for e.g. `signal_name`, `isidentifier`, or `get_distribution`) whereas failed trajectories, by contrast, rely on broad domain keywords (e.g., `default`, `pip`, `load`) that often floods the context window with irrelevant matches. The ability of the model to determine these keywords also depends on the issue specificity as it is harder to determine these keywords from under-specified/ambiguous issues, and also on the nature of the issue; bug reports often include more low-level details using error logs/code snippets for reproducing/describing the error whereas feature requests are often described at a higher-level that require adding new code in the repository. A related error behaviour that further distinguishes the two groups lies in the model's ability to determine the correct directories to search within. Successful trajectories always reach the ground truth directory as expected, whereas failed trajectories, in contrast, reach the ground truth directory in much fewer cases (often due to imprecise keyword-search). Surprisingly, we also observe some failed trajectories where the ground truth file *was viewed* during exploration, yet the model still selected a different file, indicating a failure in final-stage reasoning rather than localization.

Architectural layer selection: We observe a consistent pattern of failures in architectural layer selection that reveals the ability of the model to understand the codebase structure. For example, selecting test files instead of source code files, or utility files and classes instead of the core implementation modules, especially when these files have similar names. Successful trajectories correctly identify the implementation layer where the fix must belong, whereas failed trajectories often make mistakes here targeting the wrong layers like utility scripts, configuration files, tests, etc. instead of the actual source code that needs to be fixed.

Multi-file localization: As previously discussed, successful localization trajectories often involve identifying fewer files on average as compared to the failed trajectories. One reason for this behaviour is that our models are not explicitly trained to perform parallel tool-calling to explore multiple files simultaneously. As a result, for issues that require changes to several files, the models sometimes fail to capture cross-file dependencies and focus on a subset of files leading to incomplete localization. Another reason for this behaviour also stems from the choice of our training dataset (SWE-Smith) which is synthetically curated and edits fewer files in general.

G Additional discussion on design of reward function

This section discusses our rationale behind the design of the reward function for our training experiments. The reward function from Equation 1 reflects the ability of the LLM agent to localize relevant files, modules, and functions, and is used for all CodeScout models. We need the additional auxiliary binary reward $r^{\text{turn}}(\tau, k)$ for the 14B model because our initial experiments without this reward observed a collapse in overall reward to ≈ 0 during later training stages. We do not observe any such problems when training the 4B or 1.7B models, so we do not use this auxiliary binary reward for their training runs. We also explored auxiliary rewards to incentivize parallel tool-calling, motivated by Pan et al. [22], but found these interventions to hurt overall performance. Thus, we instead explicitly prompt the agent to use parallel tool-calling (Appendix A).

A natural concern regarding the design of $r^{\text{turn}}(\tau, k)$ is that why does it strictly assign a reward of 1 for using exactly k turns instead of flexibly assigning a reward of 1 whenever the agent terminates using $<= k$ turns. We investigate this by training the 14B model using this flexible auxiliary binary reward, $r^{\text{flexible}-\text{turn}}(\tau, k)$ that assigns a reward of 1 whenever the agent submits its predictions using $<= k$ steps, with k set to 4. All other experimental settings are identical to those used in the default CodeScout-14B model. We evaluate both the LLMs on all three benchmarks (SWE-Bench Verified, Pro, and Lite) using our setup from §4.2. We also report the average number of steps used by the agent during localization for each of these benchmarks.

As show in Table 9, we observe that using a flexible reward function for turn limits does not improve the localization performance

of the model. While the file-level F1 scores remain comparable for both the reward variants, we observe that using a flexible auxiliary reward for turn limits results in weaker function-level and module-level function scores for all the three benchmarks, with upto **6%** drop in module-level F1 score and utpo **5.14%** drop in function-level F1 score. The number of steps used by the agent to solve the task intuitively aligns with the respective reward designs - using a flexible reward for turn limit results in the agent using slightly fewer steps on average, whereas the model trained with the stricter turn limits uses ≈4 steps on average. Given the stronger localization performance of the stricter reward function for turn limits, our main paper describes results from the model trained with this reward function.

H How does choice of RL algorithm impact localization performance?

In §3.4, we described our choice of RL algorithm (GSPO) with modifications like no advantage standardization. A natural question is whether this particular configuration is optimal, or whether better performance can be achieved through other RL algorithms. We investigate this by training `Qwen3-4B-Instruct-2507` under a range of algorithm configurations, all sharing the same reward function which augments the reward from Equation 1 with the auxiliary binary term r^{turn}.

Comparability note. These ablation runs share the same learning rate, batch size, and number of rollouts as CodeScout-4B (§4.1), but use 4 turns (vs. 6) in its system prompt and the auxiliary turn reward $r^{\text{turn}}(\tau, k)$ with $k = 4$. All ablation runs are trained for 200 steps; and small performance differences may fall within the range of training variance. We compare all the checkpoints by evaluating them on SWE-Bench Pro using the setup from §4.2.

Recipe-level comparison. We first compare four representative critic-free policy-gradient recipes. These methods avoid the memory overhead of a separate value network, making them practical for multi-turn agentic training with long trajectories. They differ primarily in the *policy loss type* (how the importance ratio is computed and constrained) and the *loss reduction* (how per-token losses are aggregated):

- **GSPO** [44] (our configuration): sequence-level importance ratio with tighter clipping (ϵ = 3e-4/4e-4), sequence-mean loss reduction, and no advantage standardization.
- **GRPO** [28]: token-level PPO-clip loss ($\epsilon = 0.2$) with advantage standardization.
- **SAPO** [9]: replaces PPO clipping with a soft gating function that applies asymmetric temperatures to positive- and negative-advantage tokens, with advantage standardization.
- **Dr.GRPO** [17]: token-level PPO-clip loss (ϵ=0.2) with length-unbiased loss reduction, without advantage standardization.

As shown in the top group of Table 10, all four recipes achieve file-level F1 within a range of 47–55% and function-level F1 within the range of 22–25%, suggesting that our task is somewhat insensitive to the choice of RL algorithm. GSPO achieves the highest file-level F1, while Dr.GRPO leads on module- and function-level F1 score.

Single-factor ablations. We next isolate the effect of individual design choices by changing exactly one factor from the GSPO configuration.

Loss reduction. Standard GRPO normalizes the per-sequence loss by the response length, which introduces a length bias [17]. Dr.GRPO addresses this by dividing the per-sequence token sum by a fixed constant (the generation budget) instead. Inspired by the adoption of this technique by Pan et al. [22], we test whether it benefits GSPO by switching from seq_mean to seq_mean_token_sum_norm while keeping all other settings unchanged. This causes a substantial drop in file-level F1 (54.83% → 42.02%), suggesting it does not transfer well to GSPO's sequence-level importance ratio.

Advantage normalization. GRPO and SAPO both standardize advantages by their standard deviation, while our variant of GSPO does not. Enabling this by setting `grpo_norm_by_std=true` slightly lowers file-level F1 (54.83% → 52.73%) but improves module- and function-level metrics (31.43% → **34.37%** and 23.29% → **26.41%**).

These findings indicate that the GSPO configuration used by CodeScout performs competitively across all granularities, supporting its use as a reasonable default. More broadly, the relatively narrow performance range across recipes indicates that, for the code localization task with `Qwen3-4B/14B`, the choice of reward design (§3.3) and agent scaffold (§3.2) likely matters more than the specific RL algorithm, consistent with our recipe-oriented perspective. Finally, note that none of these approaches outperform the default recipe used to train CodeScout-4B models implying that reducing the step count from 6 to 4 hurts model performance, and it may be the case the training the 14B and 1.7B models with higher step limits may result in stronger performance.

I Example Trajectories for CodeScout-14B and CodeScout-4B

This section provides some example trajectories sampled from CodeScout-14B and CodeScout-4B models using our evaluation setup (§4.2). We use an identical GitHub issue from the SWE-Bench Verified benchmark [4] `django__django-13363` for both these models to allow a direct comparison of their tool-use behaviors and problem-solving approaches. Note that both models achieve perfectly correct localization for this instance, i.e. they achieve an F1 score of 1.0 for all three granularities: file, module, and function. Note that we omit the system prompt and user prompt (which mentions the issue description) for these example rollouts. Figures 12 and 13 present the trajectory sampled from CodeScout-14B, and figures 15 and 15 present the trajectory sampled from CodeScout-4B.

As noted in §6.2, CodeScout-14B relies only on two command-line utilities for code localization: `ripgrep (rg)` and `sed`. By analyzing the trajectory, we observe that the model uses `ripgrep` for efficient keyword-based search of the code repository and `sed` for reading specific line-ranges from relevant files. Therefore, by exploring the repository for *only* 3 steps and using *only* 2 command-line utilities, CodeScout-14B can determine the exact set of relevant set of files, classes, and functions that require modification to fix a given GitHub issue, highlighting the effectiveness of our approach.

Interestingly, CodeScout-4B uses a few more Unix command-line utilities for code localization: `rg`, `find`, `cat`, `xargs` and `sed`.

Table 9: Comparing localization performance and efficiency when training the 14B model with different choices of the auxiliary binary reward function for the turn-limit.

Benchmark	Reward function	File F1	Module F1	Function F1	Avg. # Steps
SWE-Bench Verified	$r^{\text{turn}}(\tau,k)+r^{\text{F1-file}}(y,y^\star)+r^{\text{F1-module}}(y,y^\star)+r^{\text{F1-func}}(y,y^\star)$	68.57	50.88	40.32	3.96
	$r^{\text{flex-turn}}(\tau,k)+r^{\text{F1-file}}(y,y^\star)+r^{\text{F1-module}}(y,y^\star)+r^{\text{F1-func}}(y,y^\star)$	70.37	46.43	37.30	3.78
SWE-Bench Pro	$r^{\text{turn}}(\tau,k)+r^{\text{F1-file}}(y,y^\star)+r^{\text{F1-module}}(y,y^\star)+r^{\text{F1-func}}(y,y^\star)$	53.63	37.13	28.74	3.98
	$r^{\text{flex-turn}}(\tau,k)+r^{\text{F1-file}}(y,y^\star)+r^{\text{F1-module}}(y,y^\star)+r^{\text{F1-func}}(y,y^\star)$	54.27	35.76	27.60	3.67
SWE-Bench Lite	$r^{\text{turn}}(\tau,k)+r^{\text{F1-file}}(y,y^\star)+r^{\text{F1-module}}(y,y^\star)+r^{\text{F1-func}}(y,y^\star)$	71.84	59.23	44.43	4.01
	$r^{\text{flex-turn}}(\tau,k)+r^{\text{F1-file}}(y,y^\star)+r^{\text{F1-module}}(y,y^\star)+r^{\text{F1-func}}(y,y^\star)$	70.92	53.22	39.29	3.62

Table 10: Ablation study on RL algorithm variants for `Qwen3-4B-Instruct-2507`, evaluated on SWE-Bench Pro.

Training Algorithm	Key Configuration	File F1	Module F1	Function F1
Recipe-level comparison				
GSPO	Seq-level ratio, tight clip (ϵ=3e-4/4e-4), seq-mean reduction, no std norm	**54.83%**	31.43%	23.29%
SAPO	Soft gating, seq-mean reduction, std norm	49.05%	29.82%	22.00%
Dr.GRPO	Token-level PPO clip (ϵ=0.2), length-unbiased reduction, no std norm	52.42%	33.58%	25.13%
GRPO	Token-level PPO clip (ϵ=0.2), seq-mean reduction, std norm	47.59%	30.89%	22.20%
Single-factor ablation from GSPO				
Length-unbiased reduction	`loss_reduction`: seq_mean → seq_mean_token_sum_norm	42.02%	24.88%	18.53%
Std normalization	`grpo_norm_by_std`: false → true	52.73%	**34.37%**	**26.41%**

The model mostly uses `rg`, `find` and `xargs` during keyword-based searches, and `sed` and `cat` for reading particular line-ranges from files. Furthermore, the bash commands issued by CodeScout-4B are more complex and often use more than one Unix utility in the same command chained together through piping (|). Here as well, CodeScout-4B performs code localization by exploring the repository for *only* 3 steps and using *only* 5 command-line utilities. For both the models, the command-line utilities used in these examples align with those used by the final training checkpoint in Figure 3.

A noticeable characteristic of CodeScout-4B is that it frequently performs parallel tool-calling unlike CodeScout-14B. We hypothesize that this is because of two reasons: (1) our training algorithm, and (2) choice of base model for these models. Firstly, our training algorithm does not reward/encourage parallel tool-calling and this is mainly mentioned through the system prompt: thus rollouts with parallel tool-calling have no advantage over those without any parallel tool-calling during training. Secondly, the base model for CodeScout-4B, `Qwen3-4B-Instruct-2507`, has strong instruction-following capabilities (stronger than `Qwen3-4B`, and possibly even better than `Qwen3-14B` which is the base model for CodeScout-14B). As a result, the simpler prompting mechanism is sufficient for our 4B model but not for the 14B model.

J Additional Comparison with Scaffolds that Predict a Ranked List of Top-K Locations

As opposed to prior approaches that a fixed number of top-K relevant locations, CodeScout flexibly allows the agent to dynamically predict a variable number of locations. This has the added advantage of not having to choose an appropriate value of K based on the domain or choice of evaluation benchmark. In particular, for our choice of evaluation benchmarks, the design choice of predicting a ranked list of top-K locations (with K generally set to 5) results in a higher recall but lower precision. Importantly, this design is particularly not suitable for RL fine-tuning of LLM agents, especially when K is larger than the number of ground truth locations as there is no way to determine the relevance of the remaining locations predicted by the agent and the reward function will generally ignore these predictions completely during training. Furthermore, Pan et al. [22] report that polluting the context of the coding agent is more detrimental than leaving some context out, as the agent is typically only a few searches away from recovering any remaining context, implying that precision may be more crucial than recall when augmenting code agents for downstream issue resolution.

Note that the choice of K will also determine the trade-off between precision and recall, and the results in §5 for these baselines are based on the corresponding default choice of K used by their scaffolds. We also include comparisons with these methods for different choices of K to understand the precision-recall tradeoff and

to compare these methods with CodeScout-4B and CodeScout-14B models. Table 11 presents the detailed results for three choices of K - 1, 3, and all (the default setting where K is set to 5). Impressively, CodeScout models outperform all prior methods that use a larger LLM across different choices of K even when dynamically predicting a variable number of locations. Furthermore, note that there is significant variance in the precision, recall, and F1 scores of the baseline methods for different choices of K implying that deciding the right value of K for the optimal performance is non-trivial and also depends on the characteristics of the GitHub issue and the benchmark/domain. Impressively, CodeScout strikes the right balance between precision and recall as a result of its scaffold design and our RL recipe that incorporates both precision and recall by computing the F1 scores.

K Prompts for Issue Resolution with Localization-Augmented Context

We provide detailed user and system prompts in the OpenHands software agent SDK across various experimental settings used to demonstrate the advantages of augmenting issue resolution agents with relevant localization context (§6.1). For the vanilla setup that is not augmented with any additional localization context, we provide the user prompt in Figure 16. Figures 17 and 18 describe the user prompt for the experimental setting where the issue resolution agent is augmented with locations retrieved by CodeScout-14B. Figures 19 and 20 describe the user prompt for the experimental setting where the issue resolution agent is augmented with oracle locations. All three settings use the default system prompt for OpenHands Software Agent SDK which is presented in Figures 21, 22, 23, 24, and 25.

Table 11: Comparison of CodeScout with methods that predict a ranked list of top-K locations for different choices of K (Return@K) on SWE-Bench Verified. The setting `all` refers to the default setting with K=5 used by these scaffolds.

Scaffold + LLM	Return@K	File			Function		
		F1	Prec.	Rec.	F1	Prec.	Rec.
CoSIL + Qwen2.5-32B-Instruct	1	57.25	64.65	51.36	29.11	42.51	22.13
	3	41.60	29.53	70.30	25.35	20.80	32.46
	all	30.77	19.34	83.50	22.11	14.85	55.38
Agentless + Qwen2.5-32B-Instruct	1	56.17	63.93	50.08	22.18	32.78	16.75
	3	41.46	30.13	66.45	23.51	20.83	26.98
	all	35.38	25.60	78.93	27.33	24.07	40.97
LocAgent + Qwen2.5-32B-Instruct	1	64.66	78.31	55.06	32.78	52.06	23.92
	3	49.34	38.54	68.54	27.87	24.60	32.14
	all	44.18	34.18	79.39	21.48	16.29	46.79
CodeScout-4B	-	<u>68.52</u>	71.53	67.74	<u>36.78</u>	40.71	35.72
CodeScout-14B	-	**68.57**	71.00	68.69	**40.32**	43.74	40.27

Action 1 (Terminal Tool)

```
rg "TruncDate" -t py
```

Observation 1

```
django/db/models/functions/__init__.py
5:    ExtractWeekDay, ExtractYear, Now, Trunc, TruncDate, TruncDay, TruncHour,
30:    'TruncDate', 'TruncDay', 'TruncHour', 'TruncMinute', 'TruncMonth',

django/db/models/functions/datetime.py
287:class TruncDate(TruncBase):
325:DateTimeField.register_lookup(TruncDate)

tests/db_functions/datetime/test_extract_trunc.py
13:    ExtractWeekDay, ExtractYear, Trunc, TruncDate, TruncDay, TruncHour,
839:           DTModel.objects.annotate(extracted=TruncDate('start_datetime')).order_by('start_datetime'),
```

Action 2 (Terminal Tool)

```
sed -n '1,500p' django/db/models/functions/datetime.py
```

Observation 2 (Trimmed)

```
from django.db.models.fields import (
    DateField, DateTimeField, DurationField, Field, IntegerField, TimeField,
)
...

class TruncBase(TimezoneMixin, Transform):
    def as_sql(self, compiler, connection):
        inner_sql, inner_params = compiler.compile(self.lhs)
        if isinstance(self.output_field, DateTimeField):
            tzname = self.get_tzname()
            sql = connection.ops.datetime_trunc_sql(self.kind, inner_sql, tzname)
...

class TruncDate(TruncBase):
    kind = 'date'
    lookup_name = 'date'
    output_field = DateField()

    def as_sql(self, compiler, connection):
        # Cast to date rather than truncate to date.
        lhs, lhs_params = compiler.compile(self.lhs)
        tzname = timezone.get_current_timezone_name() if settings.USE_TZ else None
        sql = connection.ops.datetime_cast_date_sql(lhs, tzname)
        return sql, lhs_params

class TruncTime(TruncBase):
    kind = 'time'
    lookup_name = 'time'
    output_field = TimeField()
...
```

Figure 12: Example trajectory sampled from CodeScout-14B for the task `django__django-13363` (continued on the next page).

Action 3 (Terminal Tool)

```
rg "get_current_timezone_name" -t py
```

Observation 3

```
django/db/models/functions/datetime.py
25:              tzname = timezone.get_current_timezone_name()
295:       tzname = timezone.get_current_timezone_name() if settings.USE_TZ else None
308:       tzname = timezone.get_current_timezone_name() if settings.USE_TZ else None

django/utils/timezone.py
64:def get_current_timezone_name():
```

Action 4 (Localization Finish Tool - Agent Loop Terminated)

```
[
    {
        "file": "django/db/models/functions/datetime.py",
        "class_name": "TruncDate",
        "function_name": "as_sql"
    },
    {
        "file": "django/db/models/functions/datetime.py",
        "class_name": "TruncTime",
        "function_name": "as_sql"
    }
]
```

Reward Computation

```
Ground truth locations:
    Files: django/db/models/functions/datetime.py
    Modules: django/db/models/functions/datetime.py:TruncDate, django/db/models/functions/datetime.py:TruncTime
    Functions: django/db/models/functions/datetime.py:TruncDate.as_sql, django/db/models/functions/datetime.py:TruncTime.as_sql
Rewards:
    File-level F1 score: 1.0
    Module-level F1 score: 1.0
    Function-level F1 score: 1.0
```

Figure 13: Example trajectory sampled from CodeScout-14B for the task `django__django-13363`.

Action 1 (Terminal Tool - 4 parallel calls)

```
# Command 1
find . -type f -name "*.py" | xargs rg -l TruncDate
# Command 2
find . -type f -name "*.py" | xargs rg -l TruncTime
# Command 3
rg -l TimeZone
# Command 4
rg -l TruncBase
```

Observation 1 (trimmed)

```
# [Results for Command 1]
./django/db/models/functions/datetime.py
./django/db/models/functions/__init__.py
./tests/db_functions/datetime/test_extract_trunc.py
# [Results for Command 2]
./django/db/models/functions/datetime.py
./django/db/models/functions/__init__.py
./tests/db_functions/datetime/test_extract_trunc.py
# [Results for Command 3 (truncated)]
django/db/backends/postgresql/base.py
django/templatetags/tz.py
# [Results for Command 4]
django/db/models/functions/datetime.py
```

Action 2 (Terminal Tool - 3 parallel calls)

```
# Command 1
cat django/db/models/functions/datetime.py | sed -n '1,500p'
# Command 2
cat django/db/models/functions/datetime.py | sed -n '501,1000p'
# Command 3
rg -n 'get_current_timezone_name' django/db/models/functions/datetime.py
```

Observation 2 (trimmed)

```
# [Results for Command 1 (truncated)]
...
from django.utils import timezone
class TimezoneMixin:
    def get_tzname(self):
...
class TruncDate(TruncBase):
    def as_sql(self, compiler, connection):
        tzname = timezone.get_current_timezone_name() if settings.USE_TZ else None
        sql = connection.ops.datetime_cast_date_sql(lhs, tzname)
...
# [Results for Command 2 (empty)]
# [Results for Command 3]
25:             tzname = timezone.get_current_timezone_name()
295:       tzname = timezone.get_current_timezone_name() if settings.USE_TZ else None
308:       tzname = timezone.get_current_timezone_name() if settings.USE_TZ else None
```

Figure 14: Example trajectory sampled from CodeScout-4B for the task `django__django-13363` (continued on the next page).

Action 3 (Terminal Tool - 2 parallel calls)

```
# Command 1
rg -n 'datetime_cast_date_sql' django/db/models/functions/datetime.py
# Command 2
rg -n 'datetime_cast_time_sql' django/db/models/functions/datetime.py
```

Observation 3

```
# [Results for Command 1]
296:       sql = connection.ops.datetime_cast_date_sql(lhs, tzname)

# [Results for Command 2]
309:       sql = connection.ops.datetime_cast_time_sql(lhs, tzname)
```

Action 4 (Localization Finish Tool - Agent Loop Terminated)

```
[
    {
        "file": "django/db/models/functions/datetime.py",
        "class_name": "TruncDate",
        "function_name": "as_sql"
    },
    {
        "file": "django/db/models/functions/datetime.py",
        "class_name": "TruncTime",
        "function_name": "as_sql"
    }
]
```

Reward Computation

```
Ground truth locations:
    Files: django/db/models/functions/datetime.py
    Modules: django/db/models/functions/datetime.py:TruncDate, django/db/models/functions/datetime.py:TruncTime
    Functions: django/db/models/functions/datetime.py:TruncDate.as_sql, django/db/models/functions/datetime.py:TruncTime.as_sql
Rewards:
    File-level F1 score: 1.0
    Module-level F1 score: 1.0
    Function-level F1 score: 1.0
```

Figure 15: Example trajectory sampled from CodeScout-4B for the task `django__django-13363`.

User Prompt for OpenHands Issue Resolution Agent (Vanilla Setting)

```
I have access to a python code repository in the directory {{ instance.repo_path }} . You can explore and modify files
     using the available tools. Consider the following issue description:
<issue_description>
{{ instance.problem_statement }}
</issue_description>
Can you help me implement the necessary changes to the repository so that the requirements specified in the <
     issue_description> are met? I've already taken care of all changes to any of the test files described in the <
     issue_description>. This means you DON'T have to modify the testing logic or any of the tests in any way! Also the
     development Python environment is already set up for you (i.e., all dependencies already installed), so you don't need
      to install other packages. Your task is to make the minimal changes to non-test files in the {{ instance.repo_path }}
      directory to ensure the <issue_description> is satisfied. Follow these phases to resolve the issue:
Phase 1. READING: read the problem and reword it in clearer terms
   1.1 If there are code or config snippets. Express in words any best practices or conventions in them.
   1.2 Hightlight message errors, method names, variables, file names, stack traces, and technical details.
   1.3 Explain the problem in clear terms.
   1.4 Enumerate the steps to reproduce the problem.
   1.5 Hightlight any best practices to take into account when testing and fixing the issue

Phase 2. RUNNING: install and run the tests on the repository
   2.1 Activate the environment by running ./opt/miniconda3/etc/profile.d/conda.sh ; conda activate testbed
   2.2 Follow the readme
   2.3 Install the environment and anything needed
   2.4 Iterate and figure out how to run the tests

Phase 3. EXPLORATION: find the files that are related to the problem and possible solutions
   3.1 Use `grep` to search for relevant methods, classes, keywords and error messages.
   3.2 Identify all files related to the problem statement.
   3.3 Propose the methods and files to fix the issue and explain why.
   3.4 From the possible file locations, select the most likely location to fix the issue.

Phase 4. TEST CREATION: before implementing any fix, create a script to reproduce and verify the issue.
   4.1 Look at existing test files in the repository to understand the test format/structure.
   4.2 Create a minimal reproduction script that reproduces the located issue.
   4.3 Run the reproduction script to confirm you are reproducing the issue.
   4.4 Adjust the reproduction script as necessary.

Phase 5. FIX ANALYSIS: state clearly the problem and how to fix it
   5.1 State clearly what the problem is.
   5.2 State clearly where the problem is located.
   5.3 State clearly how the test reproduces the issue.
   5.4 State clearly the best practices to take into account in the fix.
   5.5 State clearly how to fix the problem.

Phase 6. FIX IMPLEMENTATION: Edit the source code to implement your chosen solution.
   6.1 Make minimal, focused changes to fix the issue.

Phase 7. VERIFICATION: Test your implementation thoroughly.
   7.1 Run your reproduction script to verify the fix works.
   7.2 Add edge cases to your test script to ensure comprehensive coverage.
   7.3 Run existing tests related to the modified code to ensure you haven't broken anything.

8. FINAL REVIEW: Carefully re-read the problem description and compare your changes with the base commit {{ instance.
     base_commit }}.
   8.1 Ensure you've fully addressed all requirements.
   8.2 Run any tests in the repository related to:
     8.2.1 The issue you are fixing
     8.2.2 The files you modified
     8.2.3 The functions you changed
   8.3 If any tests fail, revise your implementation until all tests pass

Be thorough in your exploration, testing, and reasoning. It's fine if your thinking process is lengthy - quality and
     completeness are more important than brevity.
```

Figure 16: User prompt for OpenHands issue resolution agent in vanilla setup (§6.1).

User Prompt for OpenHands Issue Resolution Agent Augmented with Locations Retrieved by CodeScout-14B (continued)

```
I have access to a python code repository in the directory {{ instance.repo_path }} . You can explore and modify files
     using the available tools. Consider the following issue description:

<issue_description>
{{ instance.problem_statement }}
</issue_description>
A separate search subagent has already explored the repository to identify files, modules, and functions that are likely
     relevant to the issue.
<relevant_context>
{{ instance.search_results }}
</relevant_context>
Important notes about this context:
- The relevant context is intended to accelerate navigation, not replace your own judgment.
- The listed files/functions are very likely but not guaranteed to be sufficient.
- You may inspect files outside this list ONLY if needed, but you should prioritize the provided context.
- If you determine that some listed items are irrelevant or that additional files are required, state this explicitly in
     your analysis.

Can you help me implement the necessary changes to the repository so that the requirements specified in the <
     issue_description> are met?
I've already taken care of all changes to any of the test files described in the <issue_description>. This means you DON'T
     have to modify the testing logic or any of the tests in any way!
Also the development Python environment is already set up for you (i.e., all dependencies already installed), so you don't
     need to install other packages.
Your task is to make the minimal changes to non-test files in the {{ instance.repo_path }} directory to ensure the <
     issue_description> is satisfied.

Follow these phases to resolve the issue:

Phase 1. READING: read the problem and reword it in clearer terms
   1.1 If there are code or config snippets. Express in words any best practices or conventions in them.
   1.2 Hightlight message errors, method names, variables, file names, stack traces, and technical details.
   1.3 Explain the problem in clear terms.
   1.4 Enumerate the steps to reproduce the problem.
   1.5 Hightlight any best practices to take into account when testing and fixing the issue

Phase 2. RUNNING: install and run the tests on the repository
   2.1 Activate the environment by running
       ./opt/miniconda3/etc/profile.d/conda.sh ; conda activate testbed
   2.2 Follow the readme
   2.3 Install the environment and anything needed
   2.4 Iterate and figure out how to run the tests

Phase 3. CONTEXT REVIEW & CODE EXPLORATION: use the relevant context to orient yourself in the codebase.
   3.1 Review the <relevant_context> and understand why each listed file or function may be relevant.
   3.2 Prioritize and inspect the files and functions mentioned in the <relevant_context>.
   3.3 Read the relevant source code to understand its behavior.
   3.4 Identify discrepancies between current behavior and the expectations in the <issue_description>.
   3.5 Document any irrelevant items or missing files and adjust exploration as needed.

Phase 4. TEST CREATION: before implementing any fix, create a script to reproduce and verify the issue.
   4.1 Look at existing test files in the repository to understand the test format/structure.
   4.2 Create a minimal reproduction script that reproduces the located issue.
   4.3 Run the reproduction script to confirm you are reproducing the issue.
   4.4 Adjust the reproduction script as necessary.

Phase 5. FIX ANALYSIS: state clearly the problem and how to fix it
   5.1 State clearly what the problem is.
   5.2 State clearly where the problem is located.
   5.3 State clearly how the test reproduces the issue.
```

Figure 17: User prompt for OpenHands issue resolution agent augmented with locations retrieved by CodeScout-14B (§6.1).

User Prompt for OpenHands Issue Resolution Agent Augmented with Locations Retrieved by CodeScout-14B (continued)

```
   5.4 State clearly the best practices to take into account in the fix.
   5.5 State clearly how to fix the problem.

Phase 6. FIX IMPLEMENTATION: Edit the source code to implement your chosen solution.
   6.1 Make minimal, focused changes to fix the issue.

Phase 7. VERIFICATION: Test your implementation thoroughly.
   7.1 Run your reproduction script to verify the fix works.
   7.2 Add edge cases to your test script to ensure comprehensive coverage.
   7.3 Run existing tests related to the modified code to ensure you haven't broken anything.

Phase 8. FINAL REVIEW: Carefully re-read the problem description and compare your changes with the base commit {{ instance.
     base_commit }}.
   8.1 Ensure you've fully addressed all requirements.
   8.2 Run any tests in the repository related to:
     8.2.1 The issue you are fixing
     8.2.2 The files you modified
     8.2.3 The functions you changed
   8.3 If any tests fail, revise your implementation until all tests pass

Be thorough in your exploration, testing, and reasoning. It's fine if your thinking process is lengthy - quality and
     completeness are more important than brevity.
```

Figure 18: User prompt for OpenHands issue resolution agent augmented with locations retrieved by CodeScout-14B (§6.1).

User Prompt for OpenHands Issue Resolution Agent Augmented with Oracle Locations (continued)

```
I have access to a python code repository in the directory {{ instance.repo_path }} . You can explore and modify files
     using the available tools. Consider the following issue description:

<issue_description>
{{ instance.problem_statement }}
</issue_description>

An oracle has already determined the exact set of files, modules and functions that need modification to resolve the above
     issue description.

<oracle_context>
{{ instance.search_results }}
</oracle_context>

Important notes about this context:
- The oracle context is guaranteed to be exhaustive and precise for resolving the issue.
- In cases where the context doesn't mention functions or modules, it means that the changes either belong to the global
     scope of the file or involve adding new functions/modules in the given file.
- You may inspect files outside this list ONLY if needed, but you primarily should focus on the provided context as that is
      the one which requires modification to resolve the issue.
- The provided context will not include files that need to be created from scratch for resolving the issue.

Can you help me implement the necessary changes to the repository so that the requirements specified in the <
     issue_description> are met?
I've already taken care of all changes to any of the test files described in the <issue_description>. This means you DON'T
     have to modify the testing logic or any of the tests in any way! Also the development Python environment is already
     set up for you (i.e., all dependencies already installed), so you don't need to install other packages.
```

Figure 19: User prompt for OpenHands issue resolution agent augmented with oracle locations (§6.1).

User Prompt for OpenHands Issue Resolution Agent Augmented with Oracle Locations

```
Your task is to make the minimal changes to non-test files in the {{ instance.repo_path }} directory to ensure the <
     issue_description> is satisfied. Follow these phases to resolve the issue:

Phase 1. READING: read the problem and reword it in clearer terms
   1.1 If there are code or config snippets. Express in words any best practices or conventions in them.
   1.2 Hightlight message errors, method names, variables, file names, stack traces, and technical details.
   1.3 Explain the problem in clear terms.
   1.4 Enumerate the steps to reproduce the problem.
   1.5 Hightlight any best practices to take into account when testing and fixing the issue

Phase 2. RUNNING: install and run the tests on the repository
   2.1 Activate the environment by running
       ./opt/miniconda3/etc/profile.d/conda.sh ; conda activate testbed
   2.2 Follow the readme
   2.3 Install the environment and anything needed
   2.4 Iterate and figure out how to run the tests

Phase 3. CONTEXT REVIEW & CODE EXPLORATION: use the oracle context to orient yourself in the codebase.
   3.1 Review the <oracle_context> and understand why each listed file or function needs modification.
   3.2 Prioritize and inspect the files and functions mentioned in the <oracle_context>.
   3.3 Read the relevant source code to understand its behavior.
   3.4 Identify discrepancies between current behavior and the expectations in the <issue_description>.

Phase 4. TEST CREATION: before implementing any fix, create a script to reproduce and verify the issue.
   4.1 Look at existing test files in the repository to understand the test format/structure.
   4.2 Create a minimal reproduction script that reproduces the located issue.
   4.3 Run the reproduction script to confirm you are reproducing the issue.
   4.4 Adjust the reproduction script as necessary.

Phase 5. FIX ANALYSIS: state clearly the problem and how to fix it
   5.1 State clearly what the problem is.
   5.2 State clearly where the problem is located.
   5.3 State clearly how the test reproduces the issue.
   5.4 State clearly the best practices to take into account in the fix.
   5.5 State clearly how to fix the problem.

Phase 6. FIX IMPLEMENTATION: Edit the source code to implement your chosen solution.
   6.1 Make minimal, focused changes to fix the issue.

Phase 7. VERIFICATION: Test your implementation thoroughly.
   7.1 Run your reproduction script to verify the fix works.
   7.2 Add edge cases to your test script to ensure comprehensive coverage.
   7.3 Run existing tests related to the modified code to ensure you haven't broken anything.

Phase 8. FINAL REVIEW: Carefully re-read the problem description and compare your changes with the base commit {{ instance.
     base_commit }}.
   8.1 Ensure you've fully addressed all requirements.
   8.2 Run any tests in the repository related to:
     8.2.1 The issue you are fixing
     8.2.2 The files you modified
     8.2.3 The functions you changed
   8.3 If any tests fail, revise your implementation until all tests pass

Be thorough in your exploration, testing, and reasoning. It's fine if your thinking process is lengthy - quality and
     completeness are more important than brevity.
```

Figure 20: User prompt for OpenHands issue resolution agent augmented with oracle locations (§6.1).

Default OpenHands System Prompt used for Issue Resolution

```
You are OpenHands agent, a helpful AI assistant that can interact with a computer to solve tasks.

<ROLE>
* Your primary role is to assist users by executing commands, modifying code, and solving technical problems
    effectively. You should be thorough, methodical, and prioritize quality over speed.
* If the user asks a question, like "why is X happening", don't try to fix the problem. Just give an answer to the
    question.
</ROLE>

<MEMORY>
* Use `AGENTS.md` under the repository root as your persistent memory for repository-specific knowledge and context.
* Add important insights, patterns, and learnings to this file to improve future task performance.
* This repository skill is automatically loaded for every conversation and helps maintain context across sessions.
* For more information about skills, see: https://docs.openhands.dev/overview/skills
</MEMORY>

<EFFICIENCY>
* Each action you take is somewhat expensive. Wherever possible, combine multiple actions into a single action, e.g.
    combine multiple bash commands into one, using sed and grep to edit/view multiple files at once.
* When exploring the codebase, use efficient tools like find, grep, and git commands with appropriate filters to
    minimize unnecessary operations.
</EFFICIENCY>

<FILE_SYSTEM_GUIDELINES>
* When a user provides a file path, do NOT assume it's relative to the current working directory. First explore the
    file system to locate the file before working on it.
* If asked to edit a file, edit the file directly, rather than creating a new file with a different filename.
* For global search-and-replace operations, consider using `sed` instead of opening file editors multiple times.
* NEVER create multiple versions of the same file with different suffixes (e.g., file_test.py, file_fix.py,
    file_simple.py). Instead:
  - Always modify the original file directly when making changes
  - If you need to create a temporary file for testing, delete it once you've confirmed your solution works
  - If you decide a file you created is no longer useful, delete it instead of creating a new version
* Do NOT include documentation files explaining your changes in version control unless the user explicitly requests it
* When reproducing bugs or implementing fixes, use a single file rather than creating multiple files with different
    versions
</FILE_SYSTEM_GUIDELINES>
```

Figure 21: System prompt of OpenHands used for issue resolution experiments in §6.1.

Default OpenHands System Prompt used for Issue Resolution

```
<CODE_QUALITY>
* Write clean, efficient code with minimal comments. Avoid redundancy in comments: Do not repeat information that can
    be easily inferred from the code itself.
* When implementing solutions, focus on making the minimal changes needed to solve the problem.
* Before implementing any changes, first thoroughly understand the codebase through exploration.
* If you are adding a lot of code to a function or file, consider splitting the function or file into smaller pieces
    when appropriate.
* Place all imports at the top of the file unless explicitly requested otherwise or if placing imports at the top
    would cause issues (e.g., circular imports, conditional imports, or imports that need to be delayed for specific
    reasons).
</CODE_QUALITY>

<VERSION_CONTROL>
* If there are existing git user credentials already configured, use them and add Co-authored-by: openhands <
    openhands@all-hands.dev> to any commits messages you make. if a git config doesn't exist use \"openhands\" as the
     user.name and \"openhands@all-hands.dev\" as the user.email by default, unless explicitly instructed otherwise.
* Exercise caution with git operations. Do NOT make potentially dangerous changes (e.g., pushing to main, deleting
    repositories) unless explicitly asked to do so.
* When committing changes, use `git status` to see all modified files, and stage all files necessary for the commit.
    Use `git commit -a` whenever possible.
* Do NOT commit files that typically shouldn't go into version control (e.g., node_modules/, .env files, build
    directories, cache files, large binaries) unless explicitly instructed by the user.
* If unsure about committing certain files, check for the presence of .gitignore files or ask the user for
    clarification.
* When running git commands that may produce paged output (e.g., `git diff`, `git log`, `git show`), use `git --no-
    pager <command>` or set `GIT_PAGER=cat` to prevent the command from getting stuck waiting for interactive input.
</VERSION_CONTROL>

<PULL_REQUESTS>
* **Important**: Do not push to the remote branch and/or start a pull request unless explicitly asked to do so.
* When creating pull requests, create only ONE per session/issue unless explicitly instructed otherwise.
* When working with an existing PR, update it with new commits rather than creating additional PRs for the same issue.
* When updating a PR, preserve the original PR title and purpose, updating description only when necessary.
</PULL_REQUESTS>

<PROBLEM_SOLVING_WORKFLOW>
1. EXPLORATION: Thoroughly explore relevant files and understand the context before proposing solutions
2. ANALYSIS: Consider multiple approaches and select the most promising one
3. TESTING:
   * For bug fixes: Create tests to verify issues before implementing fixes
   * For new features: Consider test-driven development when appropriate
   * Do NOT write tests for documentation changes, README updates, configuration files, or other non-functionality
       changes
```

Figure 22: System prompt of OpenHands used for issue resolution experiments in §6.1.

Default OpenHands System Prompt used for Issue Resolution

```
   * Do not use mocks in tests unless strictly necessary and justify their use when they are used. You must always test
        real code paths in tests, NOT mocks.
   * If the repository lacks testing infrastructure and implementing tests would require extensive setup, consult with
       the user before investing time in building testing infrastructure
   * If the environment is not set up to run tests, consult with the user first before investing time to install all
       dependencies
4. IMPLEMENTATION:
   * Make focused, minimal changes to address the problem
   * Always modify existing files directly rather than creating new versions with different suffixes
   * If you create temporary files for testing, delete them after confirming your solution works
5. VERIFICATION: If the environment is set up to run tests, test your implementation thoroughly, including edge cases.
      If the environment is not set up to run tests, consult with the user first before investing time to run tests.
</PROBLEM_SOLVING_WORKFLOW>

<SELF_DOCUMENTATION>
When the user directly asks about any of the following:
- OpenHands capabilities (e.g., \"can OpenHands do...\", \"does OpenHands have...\")
- what you're able to do in second person (e.g., \"are you able...\", \"can you...\")
- how to use a specific OpenHands feature or product
- how to use the OpenHands SDK, CLI, GUI, or other OpenHands products

Get accurate information from the official OpenHands documentation at <https://docs.openhands.dev/>. The documentation
      includes:

**OpenHands SDK** (`/sdk/*`): Python library for building AI agents; Getting Started, Architecture, Guides (agent, llm,
      conversation, tools), API Reference
**OpenHands CLI** (`/openhands/usage/run-openhands/cli-mode`): Command-line interface
**OpenHands GUI** (`/openhands/usage/run-openhands/local-setup`): Local GUI and REST API
**OpenHands Cloud** (`/openhands/usage/run-openhands/cloud`): Hosted solution with integrations
**OpenHands Enterprise**: Self-hosted deployment with extended support

Always provide links to the relevant documentation pages for users who want to learn more.
</SELF_DOCUMENTATION>

<SECURITY>
# Security Policy

## OK to do without Explicit User Consent

- Download and run code from a repository specified by a user
- Open pull requests on the original repositories where the code is stored
- Install and run popular packages from pypi, npm, or other package managers
- Use APIs to work with GitHub or other platforms, unless the user asks otherwise or your task requires browsing
```

Figure 23: System prompt of OpenHands used for issue resolution experiments in §6.1.

Default OpenHands System Prompt used for Issue Resolution

```
## Do only with Explicit User Consent

- Upload code to anywhere other than the location where it was obtained from
- Upload API keys or tokens anywhere, except when using them to authenticate with the appropriate service

## Never Do

- Never perform any illegal activities, such as circumventing security to access a system that is not under your
    control or performing denial-of-service attacks on external servers
- Never run software to mine cryptocurrency

## General Security Guidelines

- Only use GITHUB_TOKEN and other credentials in ways the user has explicitly requested and would expect
</SECURITY>

<SECURITY_RISK_ASSESSMENT>
# Security Risk Policy
When using tools that support the security_risk parameter, assess the safety risk of your actions:

- **LOW**: Safe, read-only actions.
  - Viewing/summarizing content, reading project files, simple in-memory calculations.
- **MEDIUM**: Project-scoped edits or execution.
  - Modify user project files, run project scripts/tests, install project-local packages.
- **HIGH**: System-level or untrusted operations.
  - Changing system settings, global installs, elevated (`sudo`) commands, deleting critical files, downloading &
      executing untrusted code, or sending local secrets/data out.

**Global Rules**
- Always escalate to **HIGH** if sensitive data leaves the environment.
</SECURITY_RISK_ASSESSMENT>

<EXTERNAL_SERVICES>
* When interacting with external services like GitHub, GitLab, or Bitbucket, use their respective APIs instead of
    browser-based interactions whenever possible.
* Only resort to browser-based interactions with these services if specifically requested by the user or if the
    required operation cannot be performed via API.
</EXTERNAL_SERVICES>
```

Figure 24: System prompt of OpenHands used for issue resolution experiments in §6.1.

Default OpenHands System Prompt used for Issue Resolution

```
<ENVIRONMENT_SETUP>
* When user asks you to run an application, don't stop if the application is not installed. Instead, please install
    the application and run the command again.
* If you encounter missing dependencies:
  1. First, look around in the repository for existing dependency files (requirements.txt, pyproject.toml, package.
      json, Gemfile, etc.)
  2. If dependency files exist, use them to install all dependencies at once (e.g., `pip install -r requirements.txt`,
       `npm install`, etc.)
  3. Only install individual packages directly if no dependency files are found or if only specific packages are
      needed
* Similarly, if you encounter missing dependencies for essential tools requested by the user, install them when
    possible.
</ENVIRONMENT_SETUP>

<TROUBLESHOOTING>
* If you've made repeated attempts to solve a problem but tests still fail or the user reports it's still broken:
  1. Step back and reflect on 5-7 different possible sources of the problem
  2. Assess the likelihood of each possible cause
  3. Methodically address the most likely causes, starting with the highest probability
  4. Explain your reasoning process in your response to the user
* When you run into any major issue while executing a plan from the user, please don't try to directly work around it.
     Instead, propose a new plan and confirm with the user before proceeding.
</TROUBLESHOOTING>

<PROCESS_MANAGEMENT>
* When terminating processes:
  - Do NOT use general keywords with commands like `pkill -f server` or `pkill -f python` as this might accidentally
      kill other important servers or processes
  - Always use specific keywords that uniquely identify the target process
  - Prefer using `ps aux` to find the exact process ID (PID) first, then kill that specific PID
  - When possible, use more targeted approaches like finding the PID from a pidfile or using application-specific
      shutdown commands
</PROCESS_MANAGEMENT>
```

Figure 25: System prompt of OpenHands used for issue resolution experiments in §6.1..