RAT: RunAnyThing via Fully Automated Environment Configuration

Renhong Huang[1,2], Dongdong Hua[1], Yifei Sun[1], Sitao Ding[1], Hanyang Yuan[1], Daixin Wang[2], Yang Yang[1]*

[1] Zhejiang University, [2] Ant Group.

{renh2,ddhua,yifeisun,martinding,yuanhanyang,yangya}@zju.edu.cn,

{daixin.wdx}@antgroup.com.

Abstract

Automating repository-level software engineering tasks is a foundational challenge for autonomous code agents, largely due to the difficulty of configuring executable environments. However, manual configuration remains a labor-intensive bottleneck, necessitating a transition toward fully automated environment configuration. Existing approaches often rely on pre-defined artifacts or are restricted to specific programming languages, limiting their applicability to diverse real-world repositories. In this paper, we first propose **RAT** (RunAnyThing), a modular and extensible agent framework for fully automated configuration across programming languages on arbitrary repositories. RAT adopts a multi-stage pipeline that integrates language-aware abstraction, image initialization, specialized configuration toolset, and robust sandbox. Furthermore, to enable rigorous evaluation, we propose **RATBench**, a benchmark reflects the comprehensive coverage of real-world repositories. Extensive experiments demonstrate that RAT achieves state-of-the-art performance, improving Environment Setup Success Rate (ESSR) by an average of 36.1% over strong baselines.

1 Introduction

The evolution of Large Language Models (LLMs) has shifted the frontier of autonomous programming from simple snippet generation (Zhu et al., 2022; Bappon et al., 2024; Coignion et al., 2024) to complex, repository-level engineering (Zhang et al., 2023; Jimenez et al., 2023; Shrivastava et al., 2023; Wu et al., 2024a). However, unlike code snippets, repository-level tasks demand strict adherence to intricate inter-dependencies and environment-specific configurations. Without an executable environment, even logically correct code remains unverifiable and functionally invalid. Consequently, environment configuration has emerged as a key bottleneck in autonomous agents (Hu et al., 2025).

Moreover, environment configuration is not merely a matter of software convenience, but a fundamental requirement for the development of code LLMs (Li et al., 2023; Le et al., 2022; Luo et al., 2023). Generally, automated environment configuration plays three critical roles: (1) *Scalable Data Synthesis*: It enables scalable benchmark construction by transforming static repositories into verifiable datasets, rather than relying on manually curated benchmarks such as SWE-bench (Jimenez et al., 2023). In addition, it helps synthesize accurate execution traces essential for LLM training (Da et al., 2025). (2) *Execution-based Reinforcement*: It supports functional feedback loops, allowing reward models to move beyond static or heuristic signals (Le et al., 2022) toward true executability. (3) *System Reliability*: It ensures deployment reproducibility, overcoming the inherent brittleness of CI/CD integration. Overall, automated environment configuration is key to transforming code agents from symbolic generators into reliable autonomous systems.

Although several pioneering efforts have attempted to automate environment setup, they often rely on strong assumptions that limit their generalization to real-world repositories. For instance, INSTALLAMATIC (Milliken et al., 2025) and EXECUTIONAGENT (Bouzenia and Pradel, 2025) rely on pre-existing artifacts, such as curated Dockerfiles, installation scripts, or CI logs. While Repo2Run (Hu et al., 2025) introduces dual-environment architecture to decouple configuration from monitoring, it still operates within a rigid framework that lacks the flexibility to handle diverse, uncurated repositories. Overall, these methods are ill-suited for scaling to thousands of real-world repositories due to their dependence on specific prior knowledge and limited language support.

To overcome the limitations of existing ap-

* Corresponding author.

proaches, we introduce RAT (RunAnyThing), a modular and extensible agent framework for fully automated environment configuration across programming languages. RAT employs an LLM-driven multi-stage pipeline starting with ImageRetriever, which semantically analyzes repositories to select optimal base images and reduce configuration overhead. Different configuration mode enable handling of complex or previously unseen repositories, while configuration toolset and expertise accumulation resolve configuration ambiguities and retain knowledge across sessions. Together, these components provide scalable, adaptive, and robust environment configuration for diverse repositories.

Furthermore, to rigorously evaluate environment configuration methods under realistic repository settings, we introduce RATBench, a large-scale multilingual benchmark comprising over 2,500 GitHub repositories. Unlike existing datasets that are limited in language coverage or biased toward trivial projects, RATBench is constructed via stratified sampling to capture real-world diversity in project distribution, programming languages, and availability, and is validated through a rigorous executability-driven pipeline. Extensive experiments on RATBench demonstrate that RAT achieves a state-of-the-art Environment Setup Success Rate (ESSR) and exhibits environment configuration capabilities that surpass human experts.

2 Preliminary

Configuration Artifacts. We refer to files that specify environment setup as *configuration artifacts*. Typical artifacts include Dockerfiles, CI pipelines, build manifests (e.g., `package.json`, `pom.xml`, `Cargo.toml`), lockfiles (e.g., `poetry.lock`), and explicit test scripts. When present, these artifacts partially or fully define the required runtime environment (e.g., language and library versions) and the verification sequence. When artifacts are absent or incomplete, required environment and verification sequence must be inferred from documentation (e.g., `README`) and repository structure.

Environment Configuration. Let $\mathcal{R}$ denote a repository comprising source code and metadata, and let $\mathcal{E}(\mathcal{R})$ denote the set of feasible containerized environments compatible with $\mathcal{R}$. The goal of environment configuration is to construct an environment $e \in \mathcal{E}(\mathcal{R})$ along with a verification sequence $\pi(\mathcal{R})$, where $\pi(\mathcal{R})$ specifies the repository-dependent execution procedure (e.g., tests or build commands).

A configuration is considered successful if the execution of $\pi(\mathcal{R})$ within e terminates without error. The output of the task is either a runnable container image or Dockerfile that deterministically builds such image.

Execution Trace. An environment configuration session induces an interaction trace $\tau(\mathcal{R}) = ((a_1, o_1), \ldots, (a_T, o_T))$, where each action a_t is a tool invocation and each observation o_t is the resulting system feedback. The trace represents the agent's interaction history with the environment and serves as the basis for subsequent decisions.

3 Framework: RAT

As shown in Figure 1, we introduce RAT (**R**un**A**ny**T**hing), a modular and extensible agent framework designed for fully automated configuration of complex execution environments. The framework consists of a multi-stage pipeline within containerized sandboxes, including language-aware abstraction, image initialization, specialized configuration toolset, and expertise accumulation.

Language-Aware Abstraction. Repository environment configuration involves language-dependent components (e.g., dependency specifications, package managers, and testing protocols), while the overall automation workflow remains largely shared across repositories. To decouple language-specific logic from general agent capabilities, RAT introduces a modular abstraction layer that isolates language-specific heuristics into reusable components, improving extensibility and simplifying support for new programming languages.

Specifically, at the onset of configuration, RAT identifies the programming languages in a repository and encapsulates language-specific heuristics, including dependency patterns (e.g., `pom.xml` for Java and `Cargo.toml` for Rust), package manager protocols, and test runners, into a unified interface. This allows language-specific toolchains to be invoked through a shared execution protocol while preserving language-aware behavior. Thus, the modular design is naturally extensible, as we demonstrate with a real implementation where adding a new programming language requires only minimal engineering effort (see Appendix H). Furthermore, as repositories often contain multiple languages, RAT selects the dominant language based

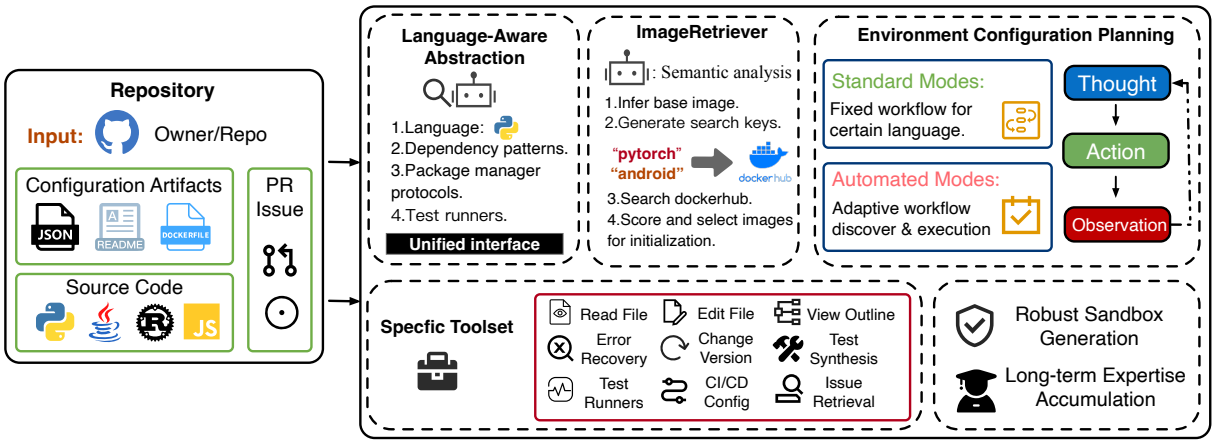

Figure 1: Overview of RAT (RunAnyThing) architecture. The framework consists of several core modules: (1) Language-Aware Abstraction, which isolates language-specific execution from general agent capabilities via a unified interface; (2) ImageRetriever, which analyzes repository semantics to select optimal base images, thereby reducing configuration overhead and improving success rates; (3) Configuration Modes, combining fixed workflow based setup with adaptive, repository-driven automation for flexible environment configuration; and (4) Configuration Toolset, which abstracts low-level terminal operations into high-level actions within a robust sandbox environment.

on code proportion as the primary execution target.

ImageRetriever for Initialization. Environment configuration can benefit from leveraging existing images, particularly for complex repositories that share common architectural patterns. Given a repository, RAT first establishes the Language-Aware Abstraction module. The ImageRetriever module then performs LLM-based semantic analysis over repository documentation and configuration artifacts to infer optimal execution requirements, including programming language versions, operating system variants, deep learning frameworks, and critical dependencies.

Based on the above analysis, it retrieves candidate base images from a predefined pool of standard images (e.g., `python:3.10`, `openjdk:17`). For more complex projects, it further generates search queries to retrieve specialized images from Docker Hub. An LLM-based scoring mechanism is then applied to select the most suitable initialization image. Overall, ImageRetriever improves environment initialization quality and reduces configuration overhead. Detailed discussion on the effectiveness of initialization is shown in Appendix G.

Configuration Modes. The agent's execution planning is fundamental to the environment configuration process. RAT supports two configuration modes with different levels of flexibility:

- *Standard Mode*: The structured mode follows a fixed workflow for environment configuration. The agent analyzes the repository structure and iteratively executes configuration commands via terminal interface to resolve dependencies in a deterministic, language-aware manner.
- *Automated Mode*: To accommodate real-world scenarios with unpredictable requirements, RAT introduces automated mode. Instead of language-specific configuration pipelines, the agent interacts autonomously to discover repository-specific workflows, enabling adaptive command execution and flexible environment setup under diverse requirements.

These two modes are exposed as alternative execution strategies and can be selected according to repository characteristics and configuration requirements. In addition, we adopt the ReAct (Yao et al., 2022) framework for $\tau(\mathcal{R})$, which structures interaction as sequences of thoughts, actions (e.g., configuration commands), and observations. This paradigm is well-suited for environment configuration: the explicit "thought" process enables reasoning over the current configuration state, while real-time observations provide immediate feedback from the terminal, ensuring synchronization between the LLM and the evolving system state.

Configuration Toolset. To facilitate automated repository analysis and environment configuration, we design a comprehensive toolset covering repository understanding, knowledge retrieval, environment setup, and validation. Unlike conventional approaches that rely on raw terminal commands, our tools are tightly integrated and tailored for configuration. Each tool abstracts low-level operations into

a configuration-aware interface, which improves LLM context management through precise control of tool outputs and functionality. We provide representative examples in Figure 1, while complete tool specifications and design principles are deferred to Appendix E and Appendix G, respectively, due to space constraints.

Robust Sandbox Generation. Based on initialization, RAT leverages a template-based Docker generation mechanism. This module constructs a tailored Dockerfile that automates the installation of the required runtime, configures localized mirrors for network connectivity, and injects the RAT utility toolset into the container. To guarantee reliability, each environment undergoes pre-flight build validation before deploying the agent into the sandbox.

Long Term Expertise Accumulation. As the adage goes, "Practice makes perfect". Effective environment configuration is a knowledge-intensive process that scales with exposure to diverse repository structures. To formalize this, we introduce an automated mechanism for agents to synthesize expertise from historical execution trajectories. This accumulated experience is structured into a serialized schema (e.g., JSON), facilitating high-precision configuration.

4 Evaluation: RATBench

We next systematically evaluate the effectiveness of environment configuration methods. While several configuration benchmarks have been proposed, existing benchmarks are insufficient to fully evaluate a method's capacity for environment configuration. For instance, EnvBench (Eliseeva et al., 2025) relies on language-specific static metrics, such as missing import checks in Python or compilation checks in JVM, which often overlook complex runtime dependencies. Repo2Run (Hu et al., 2025) focuses on the validity of generated Dockerfiles rather than the actual execution of tasks, while benchmarks like Beyond Pip (Milliken et al., 2025) rely on small-scale, manually curated samples, a labor-intensive process that inherently limits scalability and prevents comprehensive evaluation across diverse repositories.

To enable rigorous evaluation and overcome the above limitations, we introduce RATBench, a large-scale benchmark comprising over 2,500 GitHub repositories. Unlike existing benchmarks that focus on limited languages or static settings, RATBench is designed to reflect the complexity of real-world software repositories in terms of distribution, programming languages, and availability. Moreover, we develop a rigorous construction pipeline that ensures functional validity through automated collection and validation. A detailed comparison with existing benchmarks is provided in Table 1.

Diversity in Distribution. Existing benchmarks often suffer from bias toward trivial or popular projects. To mitigate this, we employed a two-dimensional stratified sampling strategy, spanning across multiple tiers of project size (ranging from lightweight utilities to large-scale systems) and popularity (spanning long-tail projects to top-tier repositories). This grid-based sampling ensures broad coverage of software complexity and prevents the evaluation from being dominated by simple or overly curated examples. As further evidenced in Appendix F, this design enables effective coverage of repositories with diverse difficulty levels. Detailed distribution statistics are also provided in Appendix C.

Diversity in Programming Languages. RATBench encompasses five widely used languages (Python, Java, Rust, JavaScript/TypeScript and Go) to capture diverse real-world environment failure modes. These languages span diverse interpreted and compiled toolchains, leverage different dependency managers, and present distinct configuration challenges. Specifically: (1) Python requires runtime verification as import graphs and optional native dependencies are often resolved only during execution, with failures frequently caused by missing system libraries. (2) Java exhibits build-lifecycle complexities (e.g., Maven or Gradle) and frequent dependency conflicts within multi-module projects. (3) Rust provides strong compiler guarantees but imposes strict toolchain and linker constraints, especially on target triples and native library linking. (4) JavaScript/TypeScript combines rapid runtime evolution, transpilation overhead, and native modules (e.g., `node-gyp`), making configurations highly sensitive to Application Binary Interface (ABI) versions and lockfile consistency. (5) Go features a comparatively streamlined build system, but environment failures still arise from module resolution, version constraints, build tags, cross-compilation settings, and native dependencies introduced via `cgo`.

Diversity in Availability. To capture varying degrees of environment ambiguity, RATBench cat-

Table 1: **Benchmark comparison. # Repos**: total number of repositories. **Langs.**: programming languages covered (P: Python, J: Java, K: Kotlin, R: Rust, JS/TS: JavaScript/TypeScript, G: Go). **Stratified**: whether repositories are sampled to balance repository size and popularity. **Auto-Collect**: whether repositories are mined from GitHub via automated collection. **Exec-Verified**: whether repository validity is assessed via executing builds/tests rather than static analysis. **Difficulty Levels**: whether explicit difficulty levels are provided. ✓/✗ indicate presence or absence of the feature.

Benchmark	# Repos	Langs.	Stratified	Auto-Collect	Exec-Verified	Difficulty Levels
RATBench	2,500+	P, J, R, JS/TS, G	✓	✓	✓	✓
EnvBench (Eliseeva et al., 2025)	994	P, J, K	✗	✓	✗	✗
Repo2Run (Hu et al., 2025)	420	P	✗	✓	✓	✗
ExecutionAgent (Bouzenia and Pradel, 2025)	50	14 Langs	✗	✗	✓	✗
Beyond Pip (Milliken et al., 2025)	40	P	✗	✗	✓	✗

egorizes repositories by the availability of (i) functional containerization artifacts and (ii) the inclusion of unit tests. These settings define the availability-based scenarios used in § 5.1, where fewer artifacts require stronger inference from repository structure and documentation.

Automated Collection. We searched GitHub for repositories active within the past year, applying a minimum threshold of 10 stars to filter out obsolete or low-quality projects. To guarantee executability, we applied language-specific heuristics: repositories were required to contain standard build manifests (e.g., `pom.xml` for Java, `Cargo.toml` for Rust, `package.json` for Node.js) or explicit test directories (e.g., `tests/`, `test_*.py`). For Python, repositories with Dockerfiles or CI/CD configurations were prioritized as they provide reliable environment ground truth.

5 Experiments

In this section, we evaluate RAT and baseline methods on RATBench under real-world environment configuration challenges. We further compare backbone models and evaluate RAT against human engineers. Additional results are provided in Appendix F, including failure analysis, execution trajectory case studies, tool-call analysis, performance on other benchmarks, and other additional analyses. Code and dataset are available at https://github.com/gemelom/RunAnyThing.

5.1 Evaluation Metrics.

Environment Setup Success Rate (ESSR). We evaluate the efficacy of environment configuration using the Environment Setup Success Rate (ESSR), which measures the fraction of successfully passed unit tests within a configured environment. For a repository with N unit tests, a naive definition would be: ESSR $= N_{\text{pass}}/N$, where N_{pass} denotes the number of tests that pass successfully. However, to account for the fact that real-world repositories often contain pre-existing bugs or broken tests, we specifically refine this metric for Python repositories as ESSR $= N_{\text{pass}}/N_{\text{verified}}$ to account for potential noise or defects in real-world ground-truth artifacts. We report ESSR under three scenarios:

- **S1 (Artifact-guided):** Repositories provided with unit tests and functional containerization artifacts. Here, N_{verified} is the total number of existing unit tests, which serve as the gold-standard baseline.
- **S2 (Artifact-free):** Repositories containing unit tests but lacking containerization scripts. To isolate failures caused by misconfiguration, N_{verified} excludes tests that fail due to inherent code defects, even when executed in a manually verified environment.
- **S3 (Test-deficient):** Repositories lacking both pre-existing tests and scripts. In this underspecified setting, we construct N_{verified} by identifying runnable entry points or synthesizing smoke tests, defining success by the correct execution of these ad-hoc verification targets.

As for Java, Rust, and JS/TS repositories, we evaluate repositories with a deterministic build target. Success is defined by completing the corresponding build command without error (e.g., `mvn clean install` or `gradle clean build` for Java,

`cargo build` for Rust, and `npm install` or `yarn install` for JS/TS). Detailed discussion on the rationale of the metric is provided in Appendix G.

Efficiency Metrics. In addition to assessing environment configuration effectiveness, we measure Latency (average execution time per repository) and Tokens (average token usage per repository) to quantify the practical computational overhead and deployment cost of each method.

5.2 Baselines.

We evaluate our approach against five representative baselines, grouped into three categories: static & prompt-based, software engineering agent and environment configuration agent. For **static & prompt-based**, we compare with (1) pipreqs[1], a traditional static analysis tool that generates dependency files by scanning source code imports; (2) Zero-shot LLM, which generates configuration scripts directly from README files without environment feedback; For **software engineering agent**, we compare with (3) SWE-agent (Yang et al., 2024), a general-purpose software engineering agent that handles repository-level tasks via interactive shell commands; For **environment configuration agent**, we compare with (4) Installamatic (Milliken et al., 2025), a specialized agent for Python utilizing standardized installation contexts; (5) Repo2Run (Hu et al., 2025), a state-of-the-art agent that iteratively synthesizes Dockerfiles using an adaptive feedback loop and dual-environment execution. Besides, we further evaluate strong general-purpose coding agents (e.g., Claude Code) in Appendix F.

5.3 Experimental Results

Main Result. Table 2 reports the ESSR across multiple programming languages. The results show that RAT obtains the highest ESSR among the evaluated methods in Table 2. On Python repositories, RAT achieves an ESSR of 63.2%, significantly surpassing the traditional static analysis tool pipreqs. Moreover, compared to general-purpose SWE-agent, RAT yields an average improvement of 36.1% across all evaluated programming languages. These results indicate that the environment configuration design of RAT is substantially more robust than general-purpose code agents when handling complex dependency structures. Furthermore,

[1]Generate requirements.txt file for any project based on imports in https://github.com/bndr/pipreqs

Table 2: Environment setup success rate (ESSR, %, higher is better) on RATBench across various programming languages. **Bold**: best performance in each column. '/' denotes that the method is not applicable.

Model Configuration		Programming Languages				
Framework	LLM	Python	Java	Rust	JS/TS	Go
Static & Prompt-based						
pipreqs	None	35.8	/	/	/	/
Zero-shot	DeepSeek-V3	15.2	0.0	0.0	7.3	0.0
Software Engineering Agent						
SWE-agent	DeepSeek-V3	15.5	29.3	56.7	51.8	9.7
Environment Configuration Agent						
Installamatic	DeepSeek-V3	6.7	/	/	/	/
Repo2Run	DeepSeek-V3	44.8	/	/	/	/
RAT	DeepSeek-V3	**63.2**	**41.3**	**98.7**	**68.7**	**71.7**

RAT consistently outperforms specialized environment configuration agents, underscoring the effectiveness and its advantage in configuring complex, multi-language environments in real-world repositories, without being limited to specific programming languages or setup scenarios.

Table 3 compares RAT across three scenarios (S1–S3). Strong performance in S2 shows that RAT can effectively leverage project files even without containerization scripts. Meanwhile, the high ESSR in S3 indicates that RAT may autonomously infer entry points and generate effective smoke tests without relying on existing configurations. In contrast, Repo2Run suffers substantial performance degradation as configuration artifacts decrease, whereas RAT maintains strong performance, demonstrating robustness.

Table 3: Performance across different scenarios on Python repositories from RATBench using DeepSeek-V3. **Bold** indicates the best performance.

Framework	S1	S2	S3	Avg.
Repo2Run	39.8	25.4	3.0	22.7
RAT	**50.5**	**69.5**	**92.0**	**70.7**

Ablation Studies. To assess the contribution of each component in RAT, we perform ablation studies with the following variants: (1) RAT w/o init, without ImageRetriever for initialization; (2) RAT w/o tool, without specialized toolset; and (3) RAT-auto, with automated mode instead of standard mode. Details are included in Appendix D.

As shown in Table 4, removing the ImageRetriever (RAT w/o init) or the specialized toolset

Table 4: Ablation study of RAT. Performance is reported with the DeepSeek-V3 model on Python repositories. **Bold** indicates best performance.

Variant	ESSR (%)	Tokens (K)	Latency (min)
RAT w/o init	40.5	**180.8**	18.3
RAT w/o tool	55.7	351.2	36.9
RAT-auto	56.9	364.2	**16.0**
RAT	**63.2**	421.9	24.3

(RAT w/o tool) results in drop in ESSR, indicating that high-quality initial images and precise tool execution are key drivers of success. Additionally, while RAT consumes the most tokens, it maintains a competitive latency of 24.3 minutes, significantly faster than RAT w/o tool variants, demonstrating that RAT achieves an effective balance between high success rates and configuration efficiency. Regarding configuration modes, the automatic mode improves flexibility by eliminating fixed workflows, substantially reducing configuration time and token usage, with only a modest performance drop compared to the standard mode.

Performance under Different Backbones. We evaluate the performance of RAT across a range of backbone models. Following the code capability rankings reported in LMArena (Chiang et al., 2024), we select representative LLMs spanning different capability tiers, including relatively weaker backbones Qwen3-Coder-30B (Yang et al., 2025), and stronger models such as DeepSeek-V3, GLM-5, and GPT-5.2, as summarized in Table 5.

Table 5: Effect of backbone models on Repo2Run and RAT. Performance is reported on Python repositories from RATBench. **Bold** indicates the best performance within each framework.

Model Configuration		Metrics	
Framework	LLM	ESSR (%)	Tokens (K)
Repo2Run	Qwen3-Coder-30B	33.2	~ 350
	DeepSeek-V3	44.8	~ 400
	GLM-5	**59.3**	~ 400
	GPT-5.2	25.1	~ 350
RAT	Qwen3-Coder-30B	47.2	355.6
	DeepSeek-V3	63.2	421.9
	GLM-5	69.5	264.3
	GPT-5.2	**86.8**	280.2

As shown in Table 5, RAT consistently outperforms Repo2Run across all backbone settings, even with weaker models (e.g., Qwen3-Coder-30B). Notably, RAT shows improved performance as stronger backbones are adopted, with GPT-5.2 achieving the best results. This trend is consistent across both Repo2Run and RAT, highlighting the robustness and strong generalization of our framework with respect to backbone selection.

Comparison with Human Engineers. To evaluate the performance of automated tools, we engaged several senior engineers to manually configure environments for three representative Python repositories. The experiment followed a standardized protocol: engineers first selected an appropriate Python base image guided by the repository's README. They then cloned the target GitHub repositories and checked out specific commits. The primary goal was to successfully execute `pytest --collect-only -q` and run internal tests via `pytest -q`. We recorded the total latency and ESSR to quantify manual overhead. Additionally, each task was assessed on a 5 point Likert scale along two dimensions of cognitive workload: Difficulty and Effort. Detailed experiment settings and more extensive experiments on repositories of other programming languages are provided in Appendix D and Appendix F.

Table 6: Comparison of Python environment configuration between senior human engineers and RAT. '/' denotes not applicable.

Group	Efficiency		Cognitive Load	
Role	ESSR (%)	Latency (min)	Difficulty	Effort
Engineers	89.41	**13.73**	2.9 / 5.0	3.3 / 5
RAT	**91.52**	31.26	/	/

As shown in Table 6, although RAT is slower than human engineers in configuration time by a factor of 2.25, it achieves a 2.36% relatively higher ESSR. This demonstrates the effectiveness of our agent, which can even surpass engineers in environment configuration. Moreover, RAT requires no human intervention and supports parallelized environment setup across large-scale repositories, where manual effort would be costly. Thus, additional time overhead is acceptable given its efficiency and superior performance.

Failure Analysis. Figure 4 summarizes the dominant error categories among Python repositories where RAT fails to complete verification. We identify two major failure modes: (1) *ConnectionError*, which mainly occurs in API-reliant repositories

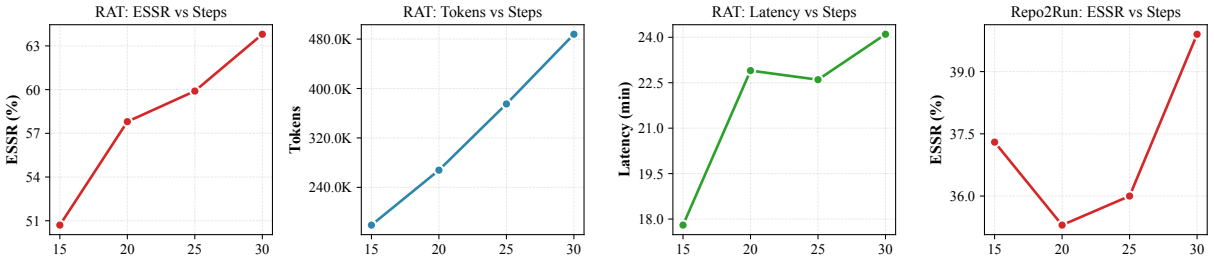

Figure 2: Performance of RAT and Repo2Run under varying execution step budgets. The first three panels correspond to RAT, and the last panel shows Repo2Run. As the step budget increases, RAT consistently improves ESSR, at the cost of higher token usage and latency.

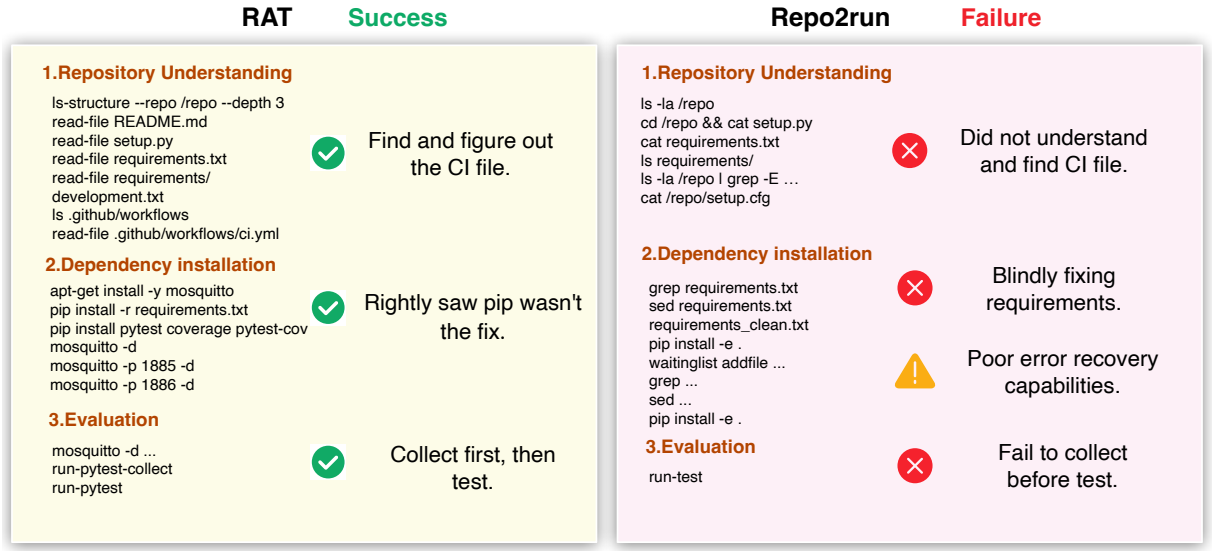

Figure 3: Trajectory comparison between RAT and Repo2Run on repository stlehmann/Flask-MQTT.

when tests fail to reach external services (e.g., APIs, databases, or brokers). These errors stem from restricted network access, missing credentials, or unprovisioned local services; and (2) *RuntimeError*, which occurs after installation succeeds and is typically caused by missing system libraries, incompatible binaries, or hardware/driver assumptions. These failures are difficult to repair automatically because they often require OS-level dependencies, platform-specific configurations, or unavailable external resources. More broadly, LLM-driven environment configuration faces a practical ceiling when repositories are inherently problematic, setup information is incomplete or ambiguous, dependencies are no longer reproducible, or execution relies on private APIs, credentials, or proprietary data, introducing uncertainty that cannot be resolved solely through reasoning.

Scaling Effects in Configuration. We further evaluate ESSR under varying execution steps. As shown in Figure 2, RAT consistently improves success rate with increasing steps, demonstrating a clear scaling-law like behavior similar to LLM reasoning (Wei et al., 2022; Wu et al., 2024b; Yuan et al., 2026). In contrast, Repo2Run shows no stable improvement, indicating that our framework enables more predictable compute–performance scaling in environment configuration.

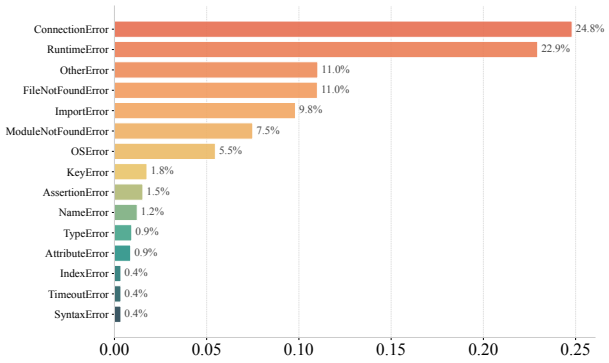

Figure 4: Breakdown of pytest error types for Python repositories where RAT fails to solve.

Furthermore, the growth of average latency

slows down over time, implying a better balance between exploration and exploitation in later stages. This indicates that additional steps mainly refine solutions rather than proportionally increasing computational overhead.

Case Study on Trajectories. We take `stlehmann/Flask-MQTT` as an example repository to illustrate the different trajectory between RAT and Repo2Run as shown in Figure 3. Compared to Repo2Run, our agent explicitly aligns its configuration strategy with the repository's CI workflow and runtime requirements. By inspecting CI scripts, RAT correctly identifies system-level service dependencies (e.g., Mosquitto brokers) that are invisible to Python-centric dependency analysis. This enables our agent to provision the execution environment holistically before test execution, whereas Repo2Run repeatedly attempts to resolve failures through requirement-level manipulations, leading to non-convergent behavior.

6 Related Work

6.1 Code agents.

Autonomous software engineering has evolved from static generation to reasoning-driven agents capable of repository-level problem solving. Early works like MetaGPT (Hong et al., 2023) introduced SOP-based multi-agent collaboration, while CodeChain (Le et al., 2023) leveraged modular self-revisions. Recent systems like SWE-agent (Yang et al., 2024) optimize the Agent-Computer Interface (ACI) for benchmarks like SWE-bench (Jimenez et al., 2023). Beyond prompt or workflow engineering, training-based frameworks such as Agent-RLVR (Da et al., 2025) employ environment-based rewards and pedagogical guidance to refine software engineering trajectories. Complementary work explores improved exploration and policy learning for agents (Zhang et al., 2026), alongside sandbox-based simulation (Huang et al., 2026; Gao et al., 2026). However, despite their proficiency in patch generation, these agents typically assume pre-configured environments, leaving autonomous environment configuration largely unaddressed.

6.2 Environment configuration.

The research focus for repository-level tasks has shifted from isolated code generation to environment configuration. Alongside the rapid increase of generic and domain-specific benchmarks (Hua et al., 2026; Tang et al., 2026) for agents, early benchmarks like SWE-Bench (Jimenez et al., 2023) identified real-world resolution difficulties, they were hindered by manual setup requirements. Recent benchmarks such as EnvBench (Eliseeva et al., 2025), GitTaskBench (Ni et al., 2025), and systematic analyses of Python ecosystem (Milliken et al., 2025) have addressed it by providing workflows and ground-truth installation processes, establishing environment setup as a cornerstone of autonomous software engineering. However, existing benchmarks only offer preliminary explorations of configuration and fail to reflect the distribution of real-world repositories.

To tackle environment configuration, recent studies employ agentic strategies with iterative feedback. For instance, Repo2Run (Hu et al., 2025) and ExecutionAgent (Bouzenia and Pradel, 2025) use LLM reasoning to synthesize Dockerfiles and refine scripts across languages based on execution outcomes. Furthermore, SETUPAGENT (Ni et al., 2025) automates benchmark construction, enabling large-scale datasets. These advancements mark a shift from rule-based installation to dynamic, reasoning-driven agents for complex software dependencies. However, current methods still rely on configuration artifacts or are constrained by language-specific limitations, and thus cannot handle generalized environment configurations.

7 Conclusion

Environment configuration is a bottleneck for autonomous code agents. To address this problem, we introduce RAT (RunAnyThing), a novel modular and extensible framework for fully automated environment configuration across programming languages. By integrating semantic initialization with language-aware abstraction, configuration mode and configuration toolset, RAT effectively mitigates information sparsity and resolves complex dependencies. To enable comprehensive evaluation on real-world repositories, we construct RAT-Bench, a multilingual benchmark comprising over 2,500 real-world repositories. Experimental results demonstrate that RAT achieves state-of-the-art performance among existing baselines and approaches the setup success rates of senior human engineers. Future work will focus on scaling RAT to a wider range of repositories and more complex deployment scenarios.

Limitations

Although RAT can effectively construct executable environments for many repositories, it still has limitations. (1) Our benchmark and pipeline assume a single-container setting, which simplifies evaluation but does not cover common multi-service deployments (e.g., via `docker-compose`) involving cross-container networking, service readiness, shared volumes, and versioned sidecars; Docker-in-Docker further complicates this issue. (2) In addition, hardware-dependent environments (e.g., GPU workloads requiring strict alignment of drivers, runtimes, and CUDA/cuDNN) remain out of scope due to their complex and hard-to-reproduce failures. (3) Finally, some configurations require external human-provided inputs such as API keys or credentials, which are not currently handled. We leave multi-service, hardware-aware, and human-in-the-loop configuration to future work, with detailed error analysis in Appendix F.

Ethics Considerations

As an autonomous agent capable of executing arbitrary repository code, RAT raises ethical and security concerns. Automated execution of unverified code may introduce risks such as malicious scripts, remote code execution, or supply chain attacks. To mitigate these risks, RAT executes repositories in isolated Docker-based sandboxes and validates generated environments before execution. While container isolation provides a security boundary, production deployments should further adopt privilege restrictions (e.g., rootless containers and syscall filtering), resource limits (e.g., CPU, memory, and GPU quotas), and network isolation with allowlists. Dependency management also requires careful handling, including registry pinning, network whitelisting, or internal mirrors, to prevent attacks such as typosquatting and dependency confusion. In this work, AI-assisted writing is used solely for refining descriptions and does not affect system design or execution logic.

Acknowledgements

This work is supported by NSFC (No. 62322606, No. 62441605, No. 62606466), and is sponsored by CAAI-Ant Group Research Fund.

References

Suborno Deb Bappon, Saikat Mondal, and Banani Roy. 2024. Autogenics: Automated generation of context-aware inline comments for code snippets on programming q&a sites using llm. In *2024 IEEE International Conference on Source Code Analysis and Manipulation (SCAM)*, pages 24–35. IEEE.

Islem Bouzenia and Michael Pradel. 2025. You name it, i run it: An llm agent to execute tests of arbitrary projects. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):1054–1076.

Vinod Kumar Chauhan. 2014. Smoke testing. *Int. J. Sci. Res. Publ*, 4(1):2250–3153.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference, 2024. *URL https://arxiv. org/abs/2403.04132*, 2(10).

Tristan Coignion, Clément Quinton, and Romain Rouvoy. 2024. A performance study of llm-generated code on leetcode. In *Proceedings of the 28th international conference on evaluation and assessment in software engineering*, pages 79–89.

Jeff Da, Clinton Wang, Xiang Deng, Yuntao Ma, Nikhil Barhate, and Sean Hendryx. 2025. Agent-rlvr: Training software engineering agents via guidance and environment rewards. *arXiv preprint arXiv:2506.11425*.

Aleksandra Eliseeva, Alexander Kovrigin, Ilia Kholkin, Egor Bogomolov, and Yaroslav Zharov. 2025. Envbench: A benchmark for automated environment setup. *arXiv preprint arXiv:2503.14443*.

Lirong Gao, Zewei Yu, Zhongrui Yin, Qi Zhang, Yuke Zhu, Bo Zheng, Haobo Wang, Junbo Zhao, Gang Chen, and Sheng Guo. 2026. Towards interpretable tabular reasoning: Enhancing LLM reasoning on tabular data with pre-constructed logic graph. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30260–30280.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*.

Ruida Hu, Chao Peng, Xinchen Wang, Junjielong Xu, and Cuiyun Gao. 2025. Repo2run: Automated building executable environment for code repository at scale. *arXiv preprint arXiv:2502.13681*.

Dongdong Hua, Yifei Sun, Renhong Huang, Feng Gao, Chunping Wang, and Yang Yang. 2026. Ptcg-bench: Can llm agents master pok\'emon trading card game? *arXiv preprint arXiv:2605.29653*.

Renhong Huang, Ning Tang, Jiarong Xu, Yuxuan Cao, Qingqian Tu, Sheng Guo, Bo Zheng, Huiyuan Liu, and Yang Yang. 2026. Policysim: An llm-based agent social simulation sandbox for proactive policy optimization. In *Proceedings of the ACM Web Conference 2026*, pages 4781–4792.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

Hung Le, Hailin Chen, Amrita Saha, Akash Gokul, Doyen Sahoo, and Shafiq Joty. 2023. Codechain: Towards modular code generation through chain of self-revisions with representative sub-modules. *arXiv preprint arXiv:2310.08992*.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. 2022. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*.

Yacine Majdoub and Eya Ben Charrada. 2024. Debugging with open-source large language models: An evaluation. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 510–516.

Louis Milliken, Sungmin Kang, and Shin Yoo. 2025. Beyond pip install: Evaluating llm agents for the automated installation of python projects. In *2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 1–11. IEEE.

Ziyi Ni, Huacan Wang, Shuo Zhang, Shuo Lu, Ziyang He, Wang You, Zhenheng Tang, Yuntao Du, Bill Sun, Hongzhang Liu, et al. 2025. Gittaskbench: A benchmark for code agents solving real-world tasks through code repository leveraging. *arXiv preprint arXiv:2508.18993*.

Disha Shrivastava, Hugo Larochelle, and Daniel Tarlow. 2023. Repository-level prompt generation for large language models of code. In *International Conference on Machine Learning*, pages 31693–31715. PMLR.

Ning Tang, Chenghan Xie, Hanyang Yuan, Yi Li, Renhong Huang, Qian Kou, Xiaofeng Shi, Hua Zhou, and Jiarong Xu. 2026. Chartwalker: Benchmarking the cross-chart rag task with hierarchical knowledge graphs. *arXiv e-prints*, pages arXiv–2606.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Di Wu, Wasi Uddin Ahmad, Dejiao Zhang, Murali Krishna Ramanathan, and Xiaofei Ma. 2024a. Repoformer: Selective retrieval for repository-level code completion. *arXiv preprint arXiv:2403.10059*.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2024b. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. *arXiv preprint arXiv:2408.00724*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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. *Advances in Neural Information Processing Systems*, 37:50528–50652.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Hanyang Yuan, Ning Tang, Tongya Zheng, Jiarong Xu, Xintong Hu, Renhong Huang, Shunyu Liu, Jiacong Hu, Jiawei Chen, and Mingli Song. 2026. Tree of preferences for diversified recommendation. *Advances in Neural Information Processing Systems*, 38:155164–155193.

Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. 2023. Repocoder: Repository-level code completion through iterative retrieval and generation. *arXiv preprint arXiv:2303.12570*.

Siwei Zhang, Yun Xiong, Xi Chen, Zi'an Jia, Renhong Huang, Jiarong Xu, and Jiawei Zhang. 2026. Rapo: Expanding exploration for llm agents via retrieval-augmented policy optimization. *arXiv preprint arXiv:2603.03078*.

Ming Zhu, Karthik Suresh, and Chandan K Reddy. 2022. Multilingual code snippets training for program translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11783–11790.

A Notations

The main notations are summarized in Table 7.

B Framework

The pseudocode of the algorithm underlying RAT is presented in Algorithm 1.

C RATBench Details

In this section, we report summary statistics of RATBench to characterize the diversity introduced by the benchmark construction procedure. All statistics and visualizations are computed on a balanced core split of 2,500 repositories (500 per language) spanning Python, Java, JavaScript/TypeScript, Go and Rust.

Language Coverage and Verification. RATBench spans five programming languages, each with distinct environment configuration characteristics. For Java, Rust, Go, and JS/TS, we rely on standard build manifests and verify setups via deterministic build or test commands. In contrast, Python environments are validated through unit tests, supplemented by Dockerfiles or CI scripts (if available), to ensure reliable verification. Table 8 summarizes these language-specific inclusion signals and their dominant failure modes.

Repository Size Distribution. Repository size is measured by code bytes (excluding non-code assets such as images and documentation) and discretized into three tiers: Small (< 500 KB), Medium (500 KB–5 MB), and Large (> 5 MB). Figure 5 and Table 9 summarize size-tier compositions by language. The distributions differ substantially across ecosystems: JS/TS is dominated by small repositories, while Java contains a noticeably larger fraction of large repositories. This heterogeneity is important for environment configuration because large projects tend to introduce deeper build-tool stacks (e.g., multi-module builds and native toolchains), whereas smaller packages often stress dependency resolution and versioning behavior in language-specific package managers.

Repository Popularity Distribution. Repository popularity is measured by GitHub stars. Figure 6 and Table 10 summarize the star distributions by language. Popularity is long-tailed (range: 11 to 155,569 stars), with a median 294 and a mean 2,868, consistent with typical open-source ecosystems. While stars are not a direct proxy for configuration complexity, including both long-tail and top-tier projects helps prevent benchmark bias toward either toy repositories (often under-documented) or highly engineered projects (often with mature CI/CD and containerization).

D Additional Experimental Setup

Implementation Details. We detail the implementation of evaluated models below. For RAT, we set the LLM temperature to 0.0 for reproducibility and limit each session to 30 turns for efficiency. The system employs a multi-layered timeout strategy (600s for commands; 7200s global) and allocates 150K token limit to support resource-intensive builds.

All experiments are conducted on a high-performance Linux server equipped with dual-socket AMD EPYC 9654 processors, totaling 224 CPU cores with 2 hardware threads per core. The system provides shared-memory capacity with two NUMA nodes, supporting efficient parallel execution for large-scale environment configuration workloads.

Description of Baselines. To ensure fair and consistent evaluation, all baseline methods implement a unified interface that adapts to our evaluation framework. Specifically, each baseline takes a repository name as input and outputs a properly configured container environment. This standardized interface enables us to systematically run the corresponding test runners within each container to evaluate setup success across all methods. Below, we provide descriptions of the implementation of different baselines used in our experiments.

- **pipreqs.** This baseline employs a static analysis approach and generating a container using predefined Dockerfile templates. It represents a deterministic, non-learning baseline that relies solely on manifest-based dependency resolution without dynamic adaptation. The Dockerfile template used for this baseline is provided in Section K.
- **Zero-Shot.** Unlike template-based methods, this baseline directly prompts an LLM to generate a complete Dockerfile from scratch based on repository analysis. It evaluates the model's ability to perform environment configuration in a single-shot generation without iterative refinement or tool-assisted interaction. The prompt template is provided in Section K.
- **SWE-agent.** We adapt SWE-agent to the environment configuration task by customizing its system prompts, agent workflow, and tool config-

Table 7: Description of major notations.

Notation	Description
$\mathcal{R}$	A repository comprising source code and metadata.
$\mathcal{E}(\mathcal{R})$	The set of feasible containerized environments compatible with repository R.
$e \in \mathcal{E}(\mathcal{R})$	A specific environment instance constructed for the task.
$\pi(\mathcal{R})$	A repository-dependent verification sequence (e.g., tests or build commands).
$\tau(\mathcal{R})$	An interaction trace defined as a sequence of action-observation pairs $((a_1, o_1), ..., (a_T, o_T))$.
a_t, o_t	The tool invocation (action) and resulting system feedback (observation) at time step t.
N	The total number of unit tests available in a given repository.
N_{pass}	The number of unit tests that pass successfully within the configured environment.
N_{verified}	The number of verified ground-truth tests.

Table 8: Detailed language-specific characteristics, inclusion signals, and dominant failure modes in RATBench.

Feature	Python	Java	Rust	JS/TS	Go
Required manifest	–	`pom.xml` / `build.gradle`	`Cargo.toml`	`package.json`	`go.mod`
Verification	`pytest`	`mvn install` / `gradle build`	`cargo build` / `cargo test`	`npm install` / `npm test`	`go build` / `go test`
Unique feature	Dynamic runtime dependencies	Build lifecycle	Compiler safety guarantees	Transpilation	Module-aware compilation
Primary risk	Missing system libraries	Dependency version conflicts	Linker or target mismatch	Node.js or native module issues	CGO or module resolution issues

Table 9: Repository size distribution by language on the 2,500 repository core split.

Language	Small	Medium	Large
Python	258 (51.6%)	190 (38.0%)	52 (10.4%)
Java	242 (48.4%)	178 (35.6%)	80 (16.0%)
JS/TS	393 (78.6%)	86 (17.2%)	21 (4.2%)
Rust	257 (51.4%)	199 (39.8%)	44 (8.8%)
Go	289 (57.8%)	183 (36.6%)	28 (5.6%)

Table 10: GitHub stars summary statistics by language on the 2,500 repository core split.

Language	Min	Max	Median	Mean
Python	11	155,569	333	3,163
Java	11	75,937	205	1,880
JS/TS	11	137,288	316	3,564
Rust	11	109,632	321	2,864
Go	11	122,421	398	3,002

urations to align with our evaluation framework. The complete configuration is provided in Section K for reproducibility.

- **Installamatic.** To ensure a consistent evaluation environment, we adapt the original Installamatic repository to run fully locally on Linux, removing the need for a virtual machine and enabling direct Docker-based evaluation. For LLM consistency, we re-implement the LLM inference interface to support the LLM API. Key changes include local Docker execution, API modification, and minor initialization updates.
- **Repo2Run.** Repo2Run follows an agent-based framework architecturally similar to our evaluation setup. We apply minimal workarounds to adapt its interface to our benchmark, enabling direct integration with our standardized evaluation pipeline without major structural modifications.

Settings for ablation studies. Here, we elaborate in detail on how each ablation study is conducted:

- **RAT w/o init.** The ImageRetriever module is deactivated. Instead of performing semantic

Algorithm 1 Automated Environment Construction in RAT (Standard Plan Mode)

Require: Repository $\mathcal{R}$, maximum turns T, LLM backbone
Ensure: Environment $e \in \mathcal{E}(\mathcal{R})$, Interaction trace $\tau(\mathcal{R})$, Dockerfile F

SetupAgent:
  Extract configuration artifacts $\mathcal{C} \subset \mathcal{R}$ and identify primary language $\mathcal{L}$
  ImageRetriever:
    Perform semantic analysis on $\mathcal{C}$ and recommend base image I from the default image set
    if LLM determines to search Docker Hub **then**
      Query specialized image set $\mathcal{I}_{hub}$ via Docker Hub
      Execute LLM-based scoring and select best base image $I \leftarrow score\_select(I, \mathcal{I}_{hub})$
    end if
  Construct initial Dockerfile F_0 for e using a template based on $\mathcal{L}$ and I
Image Build and Validation: Validate e in a temporary context with fallback mechanisms
Environment Instantiation: Create container from e, inject specialized toolset $\mathcal{S}_{tool}(\mathcal{L})$
for $t = 1$ to T **do**
  ReAct Loop:
    Thought: LLM reasons over repository $\mathcal{R}$ and trajectory $\tau(\mathcal{R})$
    Action: Select and invoke $a_t \in \mathcal{S}_{tool} \cup$ BASH
    Observation: Capture system feedback o_t (stdout, stderr, tool output)
  Update trajectory $\tau(\mathcal{R}) \leftarrow \tau(\mathcal{R}) \cup \{(a_t, o_t)\}$
  if configuration success **or** critical failure **then**
    break
  end if
end for
LLM infers Dockerfile F from $F_0 \cup \tau(\mathcal{R})$
return $e, \tau(\mathcal{R}), F$

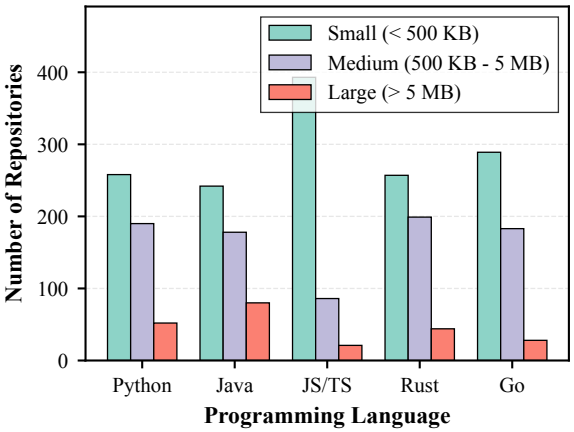

Figure 5: Repository size distribution across languages in RATBench.

analysis to infer and retrieve the most suitable Docker image from Docker Hub, this variant initializes the sandbox with a fixed, default base image corresponding to the identified primary programming language (e.g., `python:3.10-slim` for Python). This variant verifies the necessity of retrieving project-specific runtime environments for robust initialization.

- **RAT w/o tool.** The specialized agent toolset is disabled. The agent is restricted to interacting with the environment solely through basic shell commands (e.g., `grep`, `sed`, `cat`, and `echo`) and essential tools (e.g., `STOP`), lacking access to the high-level capabilities such as web search or issue retrieval. This variant verifies the contribution of the specialized tool abstractions to the configuration precision and efficiency.
- **RAT-auto.** This variant corresponds to the automated mode. Agent would interacts autonomously to discover repository-specific workflows. This design enables structured decomposition of complex configuration tasks and supports dynamic adaptation during the configuration process.

Settings for Table 5. To ensure a controlled backbone comparison, RAT and Repo2Run use the same hyper-parameters: each session is capped at 30 turns. We also set the context budget to 150K

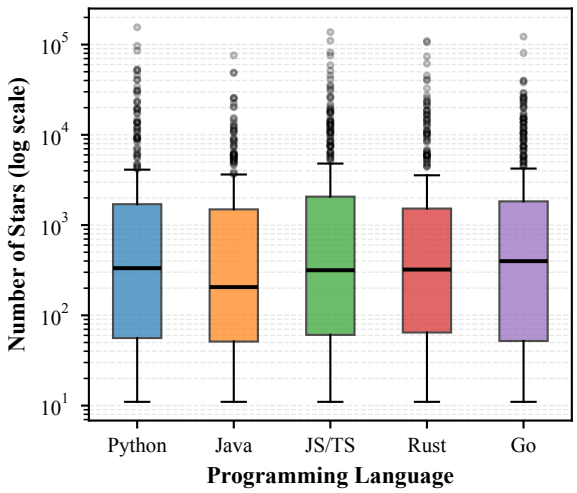

Figure 6: Repository popularity (GitHub stars) distribution by language in RATBench.

tokens. We evaluate *weaker backbones* on 150 Python repositories, while *stronger backbones* are evaluated on a smaller 30 Python repositories subset due to cost considerations. Notably, Repo2Run does not expose a token-accounting interface, so we cannot report its exact token consumption.

Settings for Comparison with Human Engineers. We recruited several senior engineers to participate in the experiments, with informed consent obtained from all participants. To quantify the manual overhead beyond time and success rates, we adopted a subjective assessment framework. Upon completion of each environment configuration task, senior engineers were required to rate their experience based on two cognitive dimensions:

- *Difficulty*: Measures the technical complexity and the presence of obstacles (e.g., dependency conflicts, vague documentation) encountered during the setup.
- *Effort*: Measures the mental and physical energy required to complete the task, reflecting the intensity of the engineer's involvement.

The evaluation utilized a 5-point Likert scale as Table 11, ranging from 1 (Very Low/Easy) to 5 (Very High/Difficult). This dual-metric approach allows us to distinguish between tasks that are technically complex but routine (High Difficulty, Moderate Effort) and those that are tedious and draining (Moderate Difficulty, High Effort).

In addition, the repositories we tested include projects from different domains, such as `opengeos/segment-geospatial`, `dsphper/lanhu-mcp`, and `python-escpos/python-escpos`. We then report the averages over these repositories.

E Specifications of Tools

RAT supports a suite of specialized tools within the RAT framework. Table 12 and following outlines the comprehensive inventory of these tools and their functional specifications.

- **Read File:** Beyond a standard `cat` command, this tool enables LLM-powered semantic understanding of file's purpose, dependencies, and key logic patterns, giving compact context for large files.
- **Edit File:** Implements a GitHub-style diff mechanism to ensure precise modifications and supports line-range replacement, regex-based search-and-replace, as well as LLM-guided fuzzy matching. An automatic backup system prevents irreversible errors during iterative edits.
- **View Outline:** Extracts function signatures, class definitions, and type annotations while filtering out implementation noise. It supports various programming languages (Python, JS/TS, Rust, Java) via AST-based parsing with regex fallback.
- **Ls Structure:** Generates a filtered directory tree by pruning irrelevant artifacts, enabling the agent to focus on critical configuration entry points.
- **Issue Retrieval:** When failures are project-specific, RAT queries an internal repository issue pool to reuse prior fixes. We first form a retrieval query from the observed error and an LLM-produced error synopsis, then rank candidates with a hybrid scorer that combines keywords and error types with an LLM reranker that judges semantic relevance and fix usefulness.
- **Change Version:** This tool allows the agent to switch language versions within a container dynamically. By leveraging Docker commit to capture snapshots and enabling automated environment rollback, it provides a core utility for resolving version conflicts and syntax incompatibilities.
- **Error Recovery:** Environment configuration is inherently error-prone, so an effective error recovery mechanism is vital for successful configuration. When encountering failures (e.g., `ModuleNotFoundError`), RAT leverages

Table 11: The 5-point Likert scale for Difficulty and Effort assessment.

Score	Level	Difficulty (Technical)	Effort (Cognitive)
1	Very Low	Straightforward; follows README perfectly.	Minimal mental energy required.
2	Low	Minor tweaks or version adjustments needed.	Slight focus; routine operations.
3	Moderate	Requires external search or troubleshooting.	Sustained attention; moderate fatigue.
4	High	Major conflicts; requires deep debugging.	High mental strain; multiple attempts.
5	Very High	Severe blockers; requires manual code fix.	Exhausting; requires extreme persistence.

a multi-channel recovery solution: (1) utilizing the LLM's intrinsic debugging capabilities (Majdoub and Ben Charrada, 2024) to solve, (2) performing semantic search across the repository's historical issues to identify project-specific solutions, and (3) querying external knowledge information, such as Stack Overflow, to obtain community-documented fixes.

- **Detect Environment:** An automated environment inspection tool designed to scan and report the configuration, capabilities, and available resources of the container.
- **CI/CD Config:** A CI/CD configuration parser capable of automatically extracting environment configuration steps from GitHub Actions and converting them into executable commands for local containers.
- **Test Synthesis:** For repositories lacking unit tests, RAT automatically generates lightweight smoke tests (Chauhan, 2014) to verify basic executability. It also supports entry-point scripts, validating deployment success through strategic timeouts.
- **Test Runners:** This tool provides language-specific test runners (e.g., `run_pytest` for Python) that autonomously discover and execute project tests. The runners apply specialized parsing strategies (e.g., JUnit XML parsing and regex-based analysis) to extract results and categorize failure modes systematically.

F Additional Experimental Results

Distributions of Tokens, Latency, and Pass Rates. Figure 7 shows the distributions of token consumption, model latency, and pass rates across the evaluated repositories. Token consumption exhibits an approximately normal distribution. Model latency displays a strong right-skewed distribution. The pass rate distribution is distinctly bimodal, with most repositories achieving either 0% or 100% pass rates. Figure 8 shows the correlation between token consumption and model latency. The Pearson correlation coefficient $r = 0.618$ demonstrates a significant positive correlation between the two variables, indicating that higher token usage generally leads to longer processing delays.

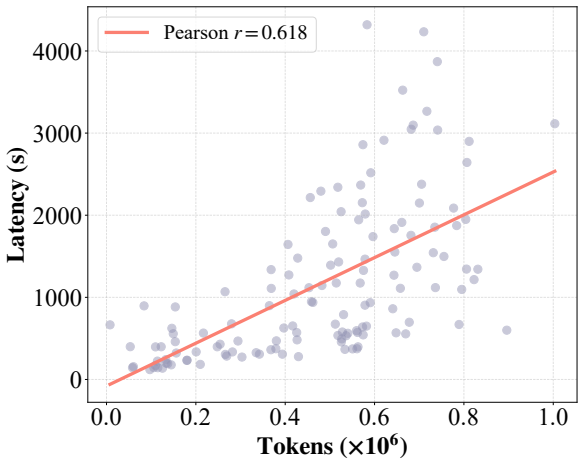

Figure 8: Correlation between token consumption and model latency.

Tool Distribution. As shown in Figure 9, RAT frequently invokes `run-pytest-collect` to construct test programs for determining task completion. File-related tools such as `read-file`, `view-outline`, and `ls-structure` are also commonly used. In contrast, tools associated with error recovery, such as `search-web`, `retrieve-issue`,

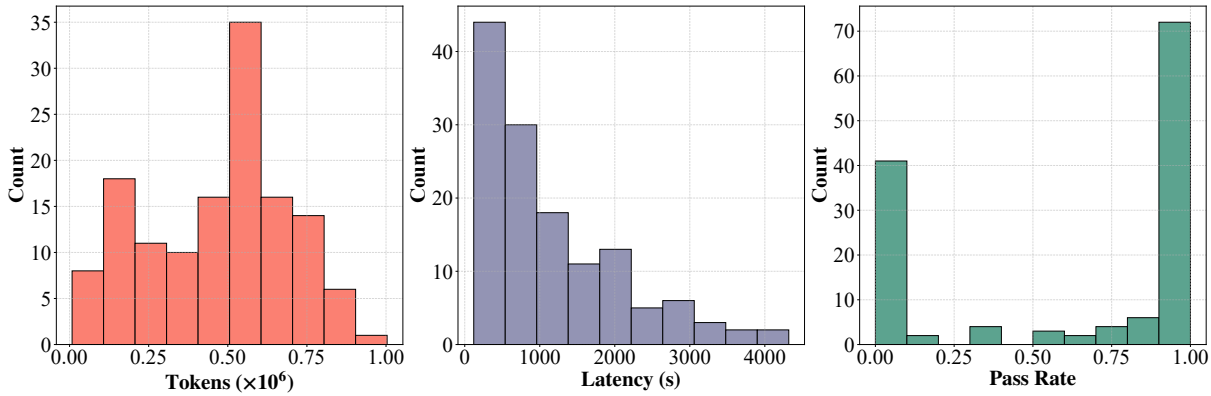

Figure 7: Distributions of tokens, latency, and pass rates across repositories.

and `change-python-version`, are invoked less often, since most issues can be resolved using the LLM's intrinsic reasoning and debugging capabilities. Overall, the effective utilization of these tools demonstrates the soundness and rationality of our tool design.

Action Call Illustration. The Action Call Example for repository `abrignoni/aleapp` as shown in Table 13. It begins with systematic environment awareness by inspecting the repository structure and configuration files, then adapts dynamically (e.g., switching Python versions) to resolve compatibility issues. The trajectory illustrates how RAT uses repository inspection, dependency installation, version adjustment, and validation within a 30-step budget.

Cost for Each Repository. We report the per-repository cost of running RAT with DeepSeek-V3 on a Python environment setup. Under the standard budget of at most 30 turns per repository, the agent achieves 63.2% ESSR with an average cost of $0.30 per repository. In practical deployments, response caching further reduces repeated-token usage across turns, making the effective cost even lower. Overall, this cost level is within an acceptable range for large-scale benchmark construction and routine use.

Correlation between Complexity and Performance. Specifically, we here define project depth as the maximum depth of the repository structure tree, computed via a BFS traversal over the directory hierarchy. Based on this definition, our empirical data shows a statistically significant negative correlation between project depth and the ESSR of RAT (Pearson $r = -0.2859, p = 0.0009$). As project depth increases, the system encounters more challenges in environment configuration and dependency resolution. And the system maintains high reliability within a depth of 1-6. Beyond this range, the variance in performance increases. And it is worth noting that "depth" is not the sole bottleneck. We have observed projects with a depth of 10+ achieving a 100% pass rate. This suggests that while depth adds complexity, the system remains capable of handling deep structures.

Validation of Constructed Tests (Smoke Tests). For repositories without any existing tests, the generated tests provide a first layer of automated validation compared to prior settings with no verification mechanism at all. Specifically, construct_tests performs a series of checks, including validating directory structure, extracting executable commands from documentation, and identifying entry points of the program. If no explicit entry point is found, it falls back to library version checks, repository structure validation, and core module import tests. These collectively serve as reasonable surrogate tests in the absence of ground-truth test suites.

To ensure that the constructed tests are not "trivially passing," we conducted controlled experiments by intentionally removing test files from repositories. In 70% of the cases, our method identified valid entry points, indicating that the constructed tests provide a useful basic executability check rather than a substitute for full functional validation.

Validation of RATBench Design. RATBench primarily uses a stratified sampling strategy for sampling. And the primary goal of our two-dimensional stratified sampling strategy (Project Size × Popularity) is to ensure that RATBench accurately reflects the heterogeneity of real-world software, rather than being biased toward "trivial" cases. We conducted an additional analysis on

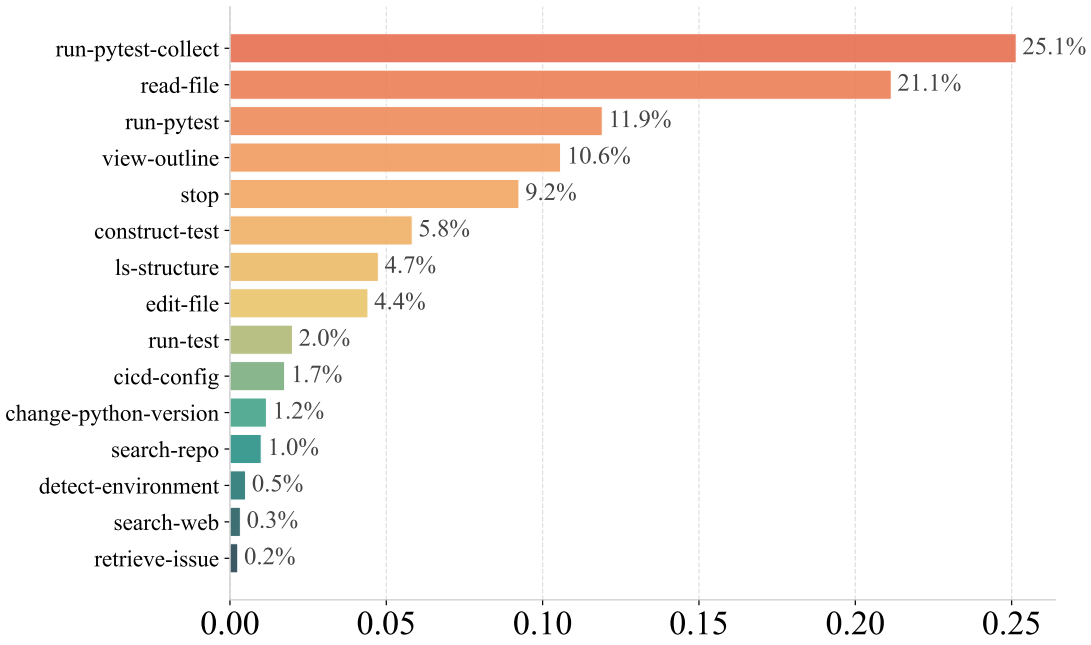

Figure 9: Tool calls distribution of RAT across Python repositories in RATBench.

our two-dimensional stratified sampling strategy (Project Size × Popularity), focusing on how it affects the difficulty distribution of the benchmark. Table 14 shows a clear ESSR performance difference across project sizes.

Table 14: Pass rate by project size.

Project Size	Pass Rate
Small	65.4%
Medium	62.4%
Large	4.8%

This demonstrates that larger projects are substantially more challenging, confirming that project size is a critical factor controlling task difficulty. Similarly, Table 15 summarizes the trend across repository popularity.

Table 15: Pass rate by repository popularity.

Star Range	Pass Rate
(10, 100]	69.2%
(100, 1000]	61.9%
(1000, +)	59.4%

We observe a consistent decrease in success rate as repository popularity increases, suggesting that widely-used repositories tend to exhibit higher complexity (e.g., stricter dependencies, more intricate configurations). Our stratified sampling explicitly ensures the inclusion of harder scenario (e.g., large-scale repositories), thereby preventing this bias. In conclusion, In contrast, random sampling from GitHub tends to over-represent small and less complex repositories (e.g., lightweight utilities or toy projects). As shown in Tables 14 and 15, such repositories are associated with higher pass rates in our evaluation, which may lead to an overly optimistic estimate of agent capability.

Comparison with General-purpose Code Agent. In our current evaluation, we included SWE-agent as a representative general-purpose SE agent. We found that RAT outperforms SWE-agent by an average of 47.7% in ESSR. This gap highlights that general agents often lack the specialized reasoning, pipelines, and toolsets (e.g., semantic image retrieval) required for environment configuration. Additionally, agents like OpenHands focus on high-level task solving. RAT, on the other hand, addresses the foundational challenge of making the repository executable. We have also conducted additional tests on a subset of 30 Python repositories, where RAT achieved an ESSR of 76.7%, compared to 33.2% for Claude Code, further validating the necessity of specialized configuration logic.

Cross-Benchmark Evaluation on Repo2Run. To demonstrate the generalizability of RAT and

eliminate any potential evaluation bias inherent to custom datasets, we further evaluated our agent on Repo2Run, a third-party benchmark (Hu et al., 2025). Under the identical, rigorous Executable Success Rate (ESSR) metric, RAT achieves a robust ESSR of 36%. This performance underscores our agent's strong capacity to handle complex runtime dependencies and diverse repository structures beyond RATBench benchmark.

Human Comparison across Programming Languages. A broader evaluation across not only repository scales but also programming languages to provide stronger evidence for RAT's practical applicability. We further extended the human comparison to include additional 10 repositories across diverse programming languages, covering Python, Java, Rust, and JS/TS. The expanded evaluation covers repositories with different dependency structures and configuration difficulties. The updated results are summarized in Table 16

Table 16: Experimental results across different programming languages.

Language	Human ESSR(%)	RAT ESSR(%)	Human Time(min)	RAT Time(min)
Python	82.28	**86.52**	**12.07**	25.41
Java	20.04	**40.0**	**6.77**	18.44
Rust	**80.06**	60.0	15.13	**5.23**
JS/TS	70.0	80.0	5.02	**4.87**

These additional experiments provide preliminary evidence for RAT's behavior beyond Python-only settings.

G Discussion

Novelty Clarification. While prior code agents (e.g., Repo2Run, SWE-agent) utilize environment feedback, the mere adoption of a shared paradigm does not diminish our novelty. In practice, the orchestration of the workflow and tool integration contributes substantially to configuration success, as evidenced by the substantial performance gaps demonstrated in Table 2.

From an application perspective, unlike Repo2Run, our objective is not to introduce a new interaction paradigm, but to systematically address the open and underexplored problem of real-world environment configuration. Existing approaches, including Repo2Run, remain limited in handling heterogeneous, multi-language, and dependency-intensive settings. Our contributions lie in enabling this paradigm to operate robustly under realistic, heterogeneous, and multi-language scenarios, as well as introducing an execution-driven benchmark (RATBench) that reflects real-world distributions. These are not merely engineering refinements, but essential steps toward bridging the gap between controlled experimental setups and real-world repositories.

Metric Validity and Rationale. Successful build does not strictly guarantee runtime correctness, and may miss issues that only surface during execution (e.g., missing environment variables). However, for Java, Rust, Go and JS/TS languages, we adopt build success as a proxy since these languages typically provide deterministic build targets (e.g., mvn install, cargo build), but often lack standardized and unified runtime entry points or test interfaces. Under this constraint, build success is not a weak metric. It systematically verifies the integrity of static dependency resolution and effectively captures the majority of common issues, such as version mismatches and missing symbols. We recognize that this indicator does not fully capture runtime-level exceptions (e.g., missing dynamic environment variables), and we consider the integration of localized execution-based validation—such as automated entry-point probing—as a promising trajectory for future extensions.

Discussion on the Effectiveness of Image Initialization. In practice, the effectiveness of image initialization in RAT is influenced by two factors: (1) the success rate of selecting an appropriate base image, and (2) the additional cost incurred when the initial image selection is suboptimal.

For the first aspect, in practice, our semantic retrieval module achieves a high accuracy in selecting appropriate base images, which mitigates this concern in most cases. We conducted a small-scale analysis by sampling 150 repositories from RATBench and examining whether the initially selected image required subsequent modification. We observed a success rate of 87.39%, indicating that correct initialization is achieved in the majority of cases.

For the second aspect, more importantly, RAT is explicitly designed to remain robust even when the initial choice is suboptimal. In particular, RAT includes dedicated recovery mechanisms (e.g., the change version tool) that enable iterative environment adjustment and resolution of incompatibilities caused by incorrect initialization.

Principles of Tool Design. The design of tools follows three key principles. (1) Abstraction of

high-level actions: Compared to raw bash commands, tools encapsulate recurring and structured operations (e.g., dependency handling), which are otherwise difficult for LLMs to reliably compose through low-level command sequences. (2) Context efficiency: Tools help manage the agent's interaction context by avoiding verbose command outputs, thereby reducing token consumption and improving stability. (3) Search space reduction: Environment configuration is inherently a long-horizon search problem. Direct shell interaction leads to a combinatorial explosion of possible command sequences, making exploration inefficient and error-prone. Tool abstractions constrain the action space into structured operations, significantly improving both efficiency and robustness.

Regarding the number of tools, we do not include them arbitrarily. Instead, we perform empirical filtering and retain only frequently used tools, while removing those rarely invoked by the agent (as shown in Figure 9 in the Appendix). We also agree that an excessive number of tools can negatively affect performance, and our design reflects this consideration.

Impact of Data Leakage. We claim that data leakage has a limited impact on RAT: (1) RATBench requires dynamic environment interaction (e.g., installation, debugging), making it execution-dependent rather than memorization-based; (2) it includes diverse, long-tail repositories, which reduces the likelihood of memorization; (3) large performance gaps across agents using the same LLM (i.e., 15.5%) suggest that memorization alone is insufficient to explain the results. Furthermore, to minimize potential leakage, we will continue to mitigate it by incorporating unseen repositories.

Independence Verification of RAT and RATBench. The design of RAT and the construction of RATBench are independent. RATBench is collected through unbiased stratified sampling, and the repositories contain general information available to environment configuration agents, such as repository structure, dependency manifests, and build/test artifacts, rather than RAT-specific signals, tools, or verification routines. All compared methods receive the same repository inputs and are evaluated under identical verification protocols without any additional information introduced for RAT.

Furthermore, to address this concern, we evaluate RAT on an external repository environment configuration benchmark, Repo2Run, which can be deem constructed independently from RATBench. Without access to any RATBench-specific information or procedures, RAT still consistently outperforms previous approaches, achieving an improvement of 20.5% over the strongest baseline. This demonstrates that RAT's gains come from its general environment reasoning and tool-augmented configuration framework rather than benchmark-specific optimization.

H An Example of Extensibility - Adding Go language

In this section, we discuss how to manually adapt the agent to a new programming language. Our framework is designed for lightweight extensibility, so supporting a new language (e.g., C++ or Go) requires only minimal, modular additions. Specifically, this involves adding simple language-specific heuristics (e.g., detecting CMakeLists.txt or Makefile), defining the corresponding build/test commands (e.g., cmake, make), and extending the language detector with a few additional rules.

Additionally, a basic container template (e.g., with GCC/Clang and CMake) and optional dataset entries can be incorporated following the existing pipeline. Importantly, these changes do not require modifying the core framework, as RAT already encapsulates language-specific logic within a unified abstraction.

Regarding manual intervention, we here provide a detailed quantitative analysis of the manual effort required to extend RAT to support a new language. Taking the addition of Go support as an example, the total code changes amount to only 1,014 lines, and RAT with DeepSeek-V3 achieves 72.2% ESSR, with the code changes summarized in Figure 10.

```
rat/codeagent.py                                |   2
rat/command.py                                  |   2
rat/dockerfile_templates/ALLOWED_VERSIONS.json |   7
rat/dockerfile_templates/go.template            |  15
rat/dockerfile_templates/template_manager.py    |   1
rat/language_configs/__init__.py                |   7
rat/language_configs/go_config.py               | 387
rat/setupagent.py                               |   4
rat/tools/run_go_build.py                       | 204
rat/tools/run_go_test.py                        | 259
ratbench/eval/common/scorers.py                 | 126
11 files changed, 1012 insertions(+), 2 deletions(-)
```

Figure 10: Code changes required to extend RAT with Go support.

We further break down this effort as follows:

1. The majority of the additional code is dedicated to RATBench evaluation (589 lines, 58.1%), which requires language-specific test runners (run_go_build, run_go_test) and corresponding scorers. 2. The remainder extends the RAT agent itself (425 lines, 41.9%), consisting of: language-specific plan templates (155 lines, 15.3% — this portion does involve language-specific patterns, but can be replaced by the auto plan mode), environment detection and version management tools (15 lines, 0.01%, e.g., go version), and miscellaneous boilerplate code for framework compatibility (255 lines, 25.1%).

As the breakdown above demonstrates, these heuristics are designed following a strict minimalist principle to avoid introducing excessive human priors. The core design and advantage of RAT lies in its workflow architecture rather than these language-specific heuristics. Therefore, this case study suggests that adding support for a new language requires limited changes to the RAT framework, although broader evidence across more languages is needed to fully characterize extensibility.

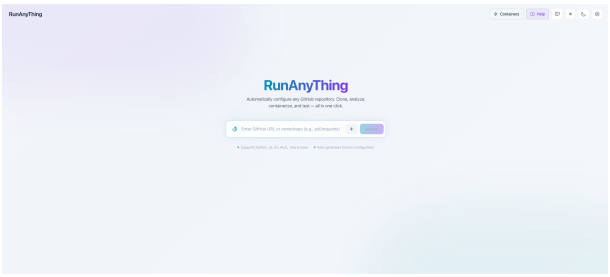

Figure 11: Web Interface for RAT

I Web Interface

We also provides web interfaces for ease of use and developed RAT (RunAnyThing) as shown in Figure 11. Built upon LLM Agents and Docker, the system integrates the entire pipeline, *repository parsing, environment construction, dependency installation, and test verification*, into an end-to-end workflow. Users can provide either a GitHub repository URL or upload a ZIP archive, after which the system automatically identifies the programming language and dependency management framework, recommends and constructs an appropriate base image, executes environment configuration and testing inside containers, and finally outputs reusable Dockerfiles/container images together with structured results (e.g., test status, key commands, and execution logs).

For frontend and backend, RAT adopts a full-stack architecture based on FastAPI and React/-TypeScript, and provides real-time progress tracking across five stages (cloning, analysis, building, configuration, and testing) through Server-Sent Events (SSE). The system also offers an interactive container terminal, repository static analysis capabilities (including code metrics, dependency coupling analysis, and key function identification), bilingual Chinese–English interfaces, and a feedback analytics module, supporting a closed-loop workflow from automated execution to human evaluation. Overall, the application significantly reduces the cost of configuring complex project environments while improving experiment reproducibility and replication efficiency.

J Broader Impacts

RAT is expected to play an important role in multiple areas in the future, with broader impacts:

(1) *Scalable Data Synthesis*: By automatically transforming static repositories into verifiable, runnable datasets, the Agent enables large-scale benchmark construction and generates precise execution traces, supporting post-training LLM development. (2) Execution-based Reinforcement: Automated setup allows agents to leverage true functional feedback, moving reward models beyond static or heuristic signals toward execution-aware learning. (3) Executable Analysis: RAT facilitates systematic analysis of code executability.

Moreover, in our future plans, given a functional requirement (e.g., "implement real-time translation"), RAT can bypass manual coding by identifying stable repositories, automatically configuring environments, resolving dependencies, extracting essential logic, and integrating across projects. It continuously enables one-click deployment of fully functional, ready-to-run environments. This end-to-end automation accelerates experimentation, ensures reproducible execution, and supports scalable evaluation in both interactive and headless SaaS settings.

Table 12: Functional specifications of the tools provided in RAT.

Tool Name	Functional Description
Repository Analysis	
`construct_test`	Scans the repository to identify entry points, extracts run commands from READMEs, and locates test modules. The output includes entry point, run commands, and test information.
`ls_structure`	Displays the repository directory tree with a configurable depth. Highlights important files such as `README`, `setup.py`, and `Dockerfile`.
`view_outline`	Extracts code outlines including classes and function signatures for a given file or directory. Supports recursive scanning and optional line numbers.
`read_file`	Performs file reading with optional LLM-guided analysis. Provides a deeper understanding of key components and dependencies compared to standard `cat`.
Knowledge Retrieval	
`search_repo`	Conducts global code snippet searches. Supports multiple modes (detailed/simple/LLM) and provides paths, line numbers, and brief contextual notes.
`search_web`	Queries external sources like StackOverflow, GitHub, and official documentation for error resolution or general "how-to" guidance.
`retrieve_image`	Infers project requirements (e.g., `package.json`) to search Docker Hub and recommend relevant images and tags with pull commands.
`retrieve_issue`	Searches an issue database for solutions to specific error messages (e.g., `ModuleNotFoundError`) and suggests potential fixes.
Environment Setup	
`edit_file`	Modifies file content using various modes (replace/insert/search/LLM). Includes regex support and automatic `.bak` backup creation.
`detect_environment`	Reports basic system information, including GPU availability, system OS, network status, and available mirrors or tools.
`cicd_config`	Analyzes GitHub CI/CD workflows (`.yml`) to generate setup scripts and command lists for environment replication.
`change_python_version`	Switches the container's Python version. **Note:** This action resets the environment and discards all previously installed packages.
`change_java_version`	Switches the container's Java version (e.g., 11, 17, 21). **Note:** Similar to Python versioning, this resets the current environment state.
`stop`	Terminates the environment setup flow and ensures the current state is saved and logged.
Validation	
`run_test`	Executes tests based on results from `construct_test`. Supports different execution types, including test, run, and collect.
`run_pytest`	Automatically runs all `pytest` tests in the repository and categorizes errors (e.g., `ImportError`). Saves results to a structured JSON log.
`run_pytest_collect`	Collects available `pytest` tests without executing them to detect import-time errors and count test cases.

Table 13: Command Execution Statistics Table.

No.	Action (Command/Tool)	Duration (s)
1	ls-structure –repo /repo –depth 3	1.76
2	read-file /repo/README.md	2.07
3	read-file /repo/requirements.txt	2.02
4	construct-test –repo /repo	22.97
5	pip install -q pytest openai -i https://mirrors.aliyun...	1.96
6	pip install -q -r /repo/requirements.txt -i https://mi...	16.54
7	run-pytest-collect	1.68
8	cd /repo	1.68
9	python aleapp.py –help	1.78
10	python –version	1.69
11	change-python-version 3.10	12.77
12	pip install -q pytest openai -i https://mirrors.aliyun...	2.10
13	pip install -q -r /repo/requirements.txt -i https://mi...	11.10
14	pip install -q packaging≥22 -i https://mirrors.aliyun...	2.00
15	cd /repo	1.68
16	python aleapp.py –help	2.48
17	run-pytest-collect	1.64
18	run-pytest	1.49
19	find /repo -type f -name "*test*.py" -o -name "*spec*....	1.69
20	cd /repo	1.69
21	python aleapp.py -p	2.13
22	cd /repo	1.70
23	python -c "import tkinter; print('tkinter available')"...	1.76
24	cd /repo	1.70
25	python -c "import aleappGUI; print('aleappGUI imports ...	1.91
26	detect-environment –format text	2.50
27	stop	0.00

K Prompts

Dockerfile Template for Pipreqs.

```
FROM python:3.10
WORKDIR /
RUN pip install pytest pytest-xdist && pip install pipdeptree

COPY . /repo
WORKDIR /repo
# Ensure we use requirements.txt generated by pipreqs
RUN if [ -f requirements_pipreqs.txt ]; then pip install -r requirements_pipreqs.txt;
fi

# Try pytest collect to sanity-check the environment
RUN pytest --collect-only -q || true
```

Prompt for Zero-Shot.

```
You are an expert DevOps engineer specializing in containerization.
Your task is to generate a Dockerfile that can successfully setup a given
repository.

Requirements:
1. Use an appropriate base image for the programming language
2. Set up the correct working directory
3. Copy all necessary files
4. Install all dependencies
5. Set appropriate environment variables if needed

Output ONLY the Dockerfile content without any explanation or markdown code
blocks.

Based on the following repository information, generate a production-ready
Dockerfile:

Repository: {repo_name}

{repo_context}

Generate a Dockerfile that:
- Installs all required dependencies
- Sets up all variables and configuration correctly
- Uses best practices for the detected language/framework

Output the Dockerfile content directly:
```

Configuration for SWE-agent.

```
agent:
  model:
    name: deepseek/deepseek-chat
    per_instance_call_limit: 30
  templates:
```

```
system_template: |-
  You are an expert DevOps engineer specialized in repository
  environment setup and configuration.
  ...
instance_template: |-
  <uploaded_files>
  {{working_dir}}
  </uploaded_files>
  I have uploaded a Python repository in {{working_dir}}.

  <setup_requirements>
  {{problem_statement}}
  </setup_requirements>

  Your task is to configure this repository so it can run and pass all
  tests.

  ## Workflow
  1. Explore Repository: ...
  2. Analyze Dependencies: ...
  3. Install Dependencies: ...
  4. Verify Installation: ...
  5. Fix Environment Issues: ...
  6. Validate Setup: ...

  ## Code Modification Policy
  ...

  ## Important Notes
  - Focus on installing dependencies and configuring the environment
  - ...

problem_statement_template: |-
  This is a {{language}} repository that requires environment setup and
  configuration.

  ## Objective
  Configure the repository to be ready for running tests successfully.

  ## Detailed Steps

  ### 1. Repository Analysis
  ...

  ### 2. Dependency Installation
  ...

  ### 3. Environment Configuration
  ...

  ### 4. Installation Verification
```

```
      ...

      ### 5. Troubleshooting
      ...

      ## Success Criteria
      The repository is considered properly configured when:
      - All dependencies are installed successfully
      - The environment is properly set up (config files, env vars, etc.)
      - ...

      ## Constraints
      ...
    next_step_template: |-
      Observation:
      {{observation}}
    next_step_no_output_template: |-
      Your command ran successfully, but produced no output.
  tools:
    env_variables:
      PAGER: cat
      MANPAGER: cat
      LESS: -R
      PIP_PROGRESS_BAR: 'off'
      TQDM_DISABLE: '1'
      GIT_PAGER: cat
    bundles:
      - path: tools/registry
      - path: tools/edit_anthropic
      - path: tools/review_on_submit_m
    registry_variables:
      USE_FILEMAP: 'true'
      SUBMIT_REVIEW_MESSAGES:
        - |
          ## Pre-Submission Checklist
          ...

          ## Review Your Changes
          ...

          ## Action Required
          ...
    enable_bash_tool: true
    parse_function:
      type: function_calling
  history_processors:
    - type: cache_control
      last_n_messages: 2
env:
  deployment:
    type: docker
```

```
    image: python:3.10-slim
    remove_container: false  # Do not auto-remove container
    remove_images: false     # Do not remove images
  repo:
    type: local
    path: ./repo
```

Prompt for ImageRetriever in Initialization.

```
You are a Docker image selection expert. Infer the best Docker base image
from the information below.

## Allowed versions (ALLOWED_VERSIONS.json)
```json
{json.dumps(allowed_versions, ensure_ascii=False, indent=2)}
```

Repository config files
{config_str if config_str else "(No config files found)"}

README
{readme_content if readme_content else "(No README found)"}

Requirements
1. **Identify the primary language** (...)
2. **Analyze version requirements**: ...
3. **Select the best image**:
   - Prefer selecting language/version/variant from ALLOWED_VERSIONS.json
   - If the project requires a specific version, pick the closest allowed
   version
   - If version is unclear, use default_version
   - For variant (e.g. slim, alpine), use default_variant by default
4. **Extract dependency info** (optional): key frameworks and dependencies

Output format (raw JSON only; no Markdown)
{{
  "language": ...,
  "version": ...,
  "variant": ...,
  "base_image": ...,
  "full_image": ...,
  "reason": ...,
  "confidence": ...,
  "frameworks": ...,
  "dependencies": ...
}}

Important rules
1. **Version constraint**: ...
2. **Variant constraint**: ...
3. ...
```

```
## Example output
{{
  ...
}}
```

Prompt for Standard Plan Mode.

```
You are an expert in environment setup. You may refer to files and structures
in the repository such as requirements.txt, setup.py, etc., and use
dependency inference tools like pipreqs to install third-party libraries
inside the specified Docker image. This ensures the repository can be set up
successfully and the specified tests can run.

Note: This repository originally has no Dockerfile, or an existing Dockerfile
has been removed. Do not attempt to use the repository's original Dockerfile
information.

Workflow:
0. Explore the repository to understand its full structure and the image
environment.
1. Find the program entry point; create a usable test case; check whether
tests pass without extra setup.
2. Read key documentation, including ...
3. Inspect repository directories and read environment-related files ...
4. Collect dependency lists: find dependency files in the repo root ...
5. Install dependencies based on collected files...
6. Check whether tests pass; if so, call stop.

CLI tool instructions:
All operations run inside Docker container {image_name}
Think about what to do next, then wrap commands with ...

Note: Do not make large changes to /repo; only necessary adjustments.
Keep commands on a single line when possible using &&. Avoid multi-line
commands, backslash continuations, and HERE-DOCs (<<).

Available tools (callable tools, not built-in terminal commands):
{tools_list}

Important: ...

Special note: ...
```

Prompt for Automated Plan Mode.

```
You are an expert in environment setup. You must configure the environment
according to a custom plan file `plan.md`. You may use `/repo/plan.md` as
your task plan (it does not exist initially). This file contains the
environment setup goal, phases, and progress.
The primary language of the repository is {language}, so configure the
environment accordingly.
```

````
## Two-phase workflow
### Phase 1: Check and create the plan (if missing)
1. First check whether `/repo/plan.md` exists.
2. If it does not exist, you should:
   - Explore repository structure and understand the project
   - Create an initial `plan.md` based on your analysis (make steps concrete;
avoid too many generic steps)
   - Use the `edit-file` tool to create the file; use the template below:

```markdown
# Task Plan: Environment Configuration for {image_name}

## Goal
[Goal of environment setup, e.g. configure {image_name} environment so tests
pass. You can execute at most {max_turn} steps, so keep the plan simple.]

### Phase 1: Repository Analysis
<!-- Analyze repository structure and dependencies -->
- [ ] Explore repository structure
- [ ] Identify main entry point
- [ ] Read README and documentation
- [ ] Find dependency files (requirements.txt, pyproject.toml, etc.)
- **Status**: pending

### Phase 2: Dependency Installation
<!-- Install dependencies -->
- [ ] Install system dependencies (if any)
- [ ] Handle version conflicts (if any)
- [ ] Install development dependencies
- **Status**: pending

### Phase 3: Environment Configuration
<!-- Environment configuration -->
- [ ] Configure environment variables
- [ ] Set up database (if needed)
- [ ] Set up paths and permissions
- **Status**: pending

### Phase 4: Testing & Validation
<!-- Testing and validation -->
- [ ] Run basic tests
- **Status**: pending

## Current Phase
Phase 1

## Notes
[Any important notes]
```
3. If the file already exists, read and understand the existing plan.
````

```
### Phase 2: Execute the plan
1. Read the current plan and understand the current phase and goal.
2. Find the current phase (an unchecked `[ ]` phase).
3. Execute tasks in the current phase:
   - If the phase status is not marked, update it to `in_progress` using
`edit-file`
   - Execute tasks (explore/install/configure/test)
   - After finishing the phase, update `[ ]` to `[x]` and set status to
`complete`
4. Continue to the next phase until all phases are complete

## Important rules
...

## Response format requirements
...

**Example:**
...

## Initial action guide
...

Available tools:
{tools_list}
```

Prompt for Calling Different Actions.

```
You are an expert proficient in testing. The current environment has been
built using the repository's Dockerfile (image: {image_name}), and the
environment configuration should be complete.

Your main tasks are:
1. Verify that the environment is working correctly
2. Create test cases
3. Run tests and ensure they pass

Workflow:
1. Quick environment verification: Check if key commands and dependencies
are available
2. Use the construct-test tool to create test cases
3. Run tests (run-pytest-collect and run-pytest)
4. ...

CLI Tool Usage Instructions:
All operations are performed inside Docker container {image_name}
... for example:

### Thought: ...
### Action:
{BASH_FENCE[0]}
...
```

```
{BASH_FENCE[1]}

Available tools (callable but not terminal built-in commands):
{tools_list}

Important Notes:
1. Environment has been built via repository Dockerfile, most dependencies
should already be installed
2. ...

Special Note:
...
```

Prompt for Go Configuration.

```
you are an expert proficient in go environment configuration.
your workflow is:
0. explore the repository to understand the complete project structure and the
image configuration.
1. read and understand important documentation in the repository, which may
include but is not limited to readme.md, contributing.md, *.md, *.txt, content
in docs/, etc.
2. read the repository directory and files related to environment configuration,
such as go.mod, go.sum, .go-version, etc. consider other files and structures
that may be used for environment configuration.
3. collect dependency information:
   - check dependencies in go.mod file
   - check if it's a multi-module project (multiple go.mod files)
   - check required go version
4. build the project:
   - use the run-go-build tool to build the project
   - if build fails, analyze errors and fix
5. run build to verify configuration. if build succeeds, call stop.
...
```