

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 CVEDRL: AN EFFICIENT CODE VERIFIER VIA DIFFICULTY-AWARE REINFORCEMENT LEARNING

Anonymous authors

Paper under double-blind review

ABSTRACT

Code verifier is the key to the post-verification process in large language model (LLM) code generation. However, supervised fine-tuning (SFT) methods suffer from dataset scarcity, high error and failure rates, and severe inference delay. In this work, we adapt reinforcement learning to train an efficient code verifier, CVeDRL, which substantially alleviates these challenges and balances performance and efficiency in only a 0.6B scale. First, we design syntax and functionality rewards and employ GRPO to train the base code verifier. However, preliminary experiments indicated that the base model could not produce effective unit tests for difficult branches and samples. Then we propose Branch-Difficulty-aware and Sample-Difficulty-aware reinforcement learning based on exponential reward shaping and static analysis metrics (Halstead Complexity and Maintainability Index). Experimental results show that CVeDRL significantly outperforms the vanilla model while remaining competitive with state-of-the-art models such as GPT-4o-mini and GPT-3.5 in pass rate, assertion failure rate, and code coverage, etc. Furthermore, CVeDRL-0.6B improves inference efficiency by more than 20x compared with LLM trained with SFT method. Code is available at <https://anonymous.4open.science/r/CVeDRL-DF1A/>

1 INTRODUCTION

Recently, large language models (LLMs) have shown impressive capabilities in code generation (Achiam et al., 2023; Dubey et al., 2024; Jaech et al., 2024; Guo et al., 2025). Although LLMs are able to quickly generate code solutions, they struggle to produce correct code in a single attempt. Researchers investigate inference-time scaling methods to alleviate this difficulty, where LLMs first employ repeated sampling to output multiple code results, and then a code verifier selects the final best result (Lightman et al., 2024b; Brown et al., 2024). Consequently, the performance of the code verifier is the key to the success of the code generation task for large language models (Cobbe et al., 2021; Lightman et al., 2024b; Liu et al., 2025; Zhao et al., 2025).

Code verifiers take the problem description and the candidate code solution as input and leverage LLMs to generate unit tests, which consist of reasonable input and corresponding output pairs. By executing candidate code solutions and their unit tests, the execution outcomes are examined from compilers and interpreters to identify the optimal code solutions among all candidates. In fact, code verifiers are specific LLMs for code generation, as unit tests are special code snippets. As a result, instruction supervised fine-tuning (SFT) can endow LLMs with unit test generation ability and make them qualified code verifiers (Ma et al., 2025). However, there are three challenges in the SFT strategy. First, large-scale high-quality SFT data is unavailable for unit test generation. Although Ma et al. (2025) presents an automatic data pipeline, a lot of incorrect unit tests exist in the final results. Second, the SFT-based model suffers from a high error rate and failure rate on generated unit tests. Besides, SFT-based code verifiers have to repeatedly sample multiple unit tests to mitigate error and failure issues, causing a serious efficiency bottleneck.

The reinforcement learning (RL) training paradigm induces and supervises LLMs to explore appropriate answers through delicate reward signals, which shows substantial potential in the LLM post-training period (Jaech et al., 2024; Shao et al., 2024; Guo et al., 2025). Training efficient code verifiers by RL is a promising approach to address the above challenges. For one thing, RL training only needs the problem description and the candidate code, without relying on corresponding unit

tests, which avoids the requirement of high-quality datasets. For another thing, common metrics such as pass rate and line coverage of unit tests in software engineering can be adapted as RL reward functions, directly reducing error and failure rates during training. In addition, unleashing the post-training capability of small-scale LLMs and decreasing repeated sampling by RL contribute to efficiency improvement.

In this work, we adapt reinforcement learning methods to train an efficient code verifier CVeDRL, which has both performance and efficiency advantages at a scale of 0.6B. To ensure that the generated unit test cases align with the formatting requirements of test suites and achieve extensive branch coverage for enhanced test quality, we propose a syntax-functionality composite reward and utilize Group Reward Policy Optimization (GRPO) Shao et al. (2024) for training. Specifically, we theoretically analyze the relationship among test-case pass rates, branch coverage, and the efficacy of the code verifier. However, preliminary experiments indicated that our initial approach inadequately distinguished boundary branches and variations in sample difficulty associated with code solutions. Consequently, we introduce a branch-difficulty-aware mechanism grounded in exponential reward shaping and sample-difficulty-aware technique with static analysis metrics to handle both problems respectively.

We perform extensive experiments to verify CVeDRL across three datasets and four policy models. Experimental results demonstrate that CVeDRL-0.6B significantly improves the performance of open-source policy models and also enhances the effectiveness of closed-source models. Additionally, to assess the intrinsic quality of generated unit tests, we directly evaluate various metrics such as error rates, pass rates, and line coverage across three datasets. CVeDRL-0.6B consistently outperforms various baseline models in the task of unit test generation, achieving a considerably higher test-case pass rate on MBPP+ (yielding a 17.55% increase compared to GPT-4o-mini) with marginally higher coverage, and substantiating our theoretical analyses. Furthermore, CVeDRL-0.6B gains more than 20x inference efficiency improvement on token throughput, compared to the traditional SFT model CodeRM.

The contributions of our work are as follows:

- We theoretically analyze the interplay among test case pass rates, branch coverage, and the performance of code verifiers.
- We propose CVeDRL, an efficient code verifier trained by GRPO with syntax and functionality rewards, alongside a novel branch-sample difficulty-aware mechanism incorporating exponential reward shaping and static analysis metrics.
- Extensive experiments demonstrate that CVeDRL-0.6B achieves sota performances on LLM code verification and unit test generation across six groups of main experiment overall and improves inference efficiency by more than 20x compared with SFT models.

2 RELATED WORK

Enhancement of code generation via unit testing. Enhancing the reliability of code generation with unit test cases can be framed as a two-phase process: improving model training and refining inference. Extensive prior work integrates unit test results into the training process to improve the accuracy of code generators Liu et al. (2023a); Dou et al. (2024). CURE Wang et al. (2025c) demonstrates that co-training tests and code yields more robust behavior in the training phase, and subsequent RL variants incorporate real execution feedback from unit test runs to further shape model policies. Prior research validates codes generated by LLM with automated testing at inference time with optimal solution selection. For instance, MBR-EXEC Shi et al. (2022b) applies Bayesian risk decoding to rerank code candidates by their pass rates. CodeRM Ma et al. (2025) leverages a distilled “test generator” model to produce targeted unit tests that guide solution selection. Building on these insights, our method combines test accuracy and line-coverage signals within a unified reward function, thereby both sharpening higher quality of unit test generation and enabling more reliable optimal code selection at inference.

Reinforcement Learning for LLMs. Recent work systematically frames text generation as a Markov decision process, adapting the classical RL pipeline to large language models by treating each token prediction as an action and defining rewards that capture response quality Wang et al.

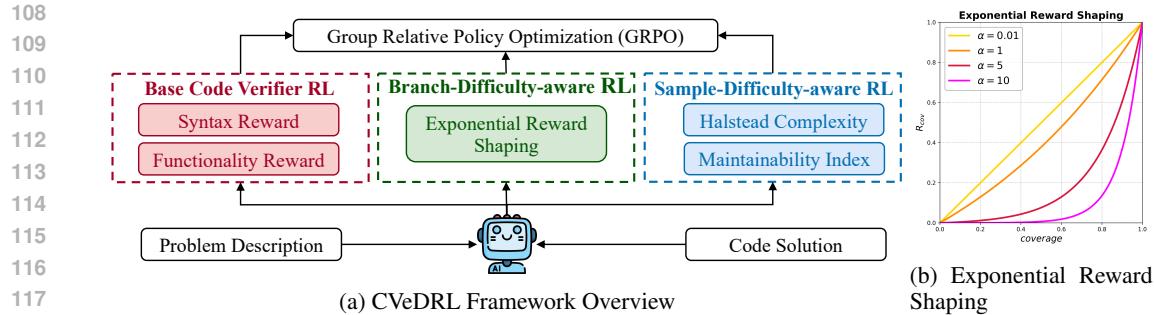


Figure 1: (a) We design syntax and functionality rewards and employ GRPO to train the base code verifier. Base model struggles to produce effective unit test cases for the difficult branches and samples. Therefore, we propose Branch-Difficulty-aware and Sample-Difficulty-aware reinforcement learning based on exponential reward shaping and static analysis metrics (Halstead Complexity and Maintainability Index). (b) Exponential reward shaping modifies the coverage reward function from a linear format into an exponential format.

(2025a). Prominent architectures extend base models (e.g., Alpaca, LLaMA) via policy-gradient, proximal policy optimization, and actor-critic algorithms Dong et al. (2025) to iteratively enhance coherence and factuality Hu et al. (2025). Two main reward-model techniques, Reinforcement Learning from Human Feedback (RLHF) Christiano et al. (2023) using curated human preferences and Reinforcement Learning from AI Feedback (RLAIF) Lee et al. (2024), have become standard for aligning outputs with desired behaviors. Direct Preference Optimization (DPO) Rafailov et al. (2024) methods bypass explicit reward functions by directly fitting model parameters to preference data, achieving comparable alignment with lower complexity. Recent extensions of DPO include Group Robust Preference Optimization (GRPO) Ramesh et al. (2024), which adaptively re-weights group-specific losses to ensure worst-case group performance under data imbalance. Balanced Preference Optimization (BPO) Wang et al. (2025b) applies dynamic reward-margin balancing to stabilize preference updates in DPO.

Unit test generation. Unit test generation automates the creation of test cases to verify code correctness, expedite bug detection, and uphold source quality. Traditional approaches often employ symbolic analysis Galeotti et al. (2013) and meta-heuristic Harman & Jones (2001); McMinn (2004); Harman et al. (2012) algorithms to craft tests. LLM-based techniques have garnered attention for their efficiency, interpretability, and readable outputs Jiang et al. (2024). Prior research treat LLM as an auxiliary part for traditional methods to help the exploration Lemieux et al. (2023) or mutation Brownlee et al. (2023). More recently, LLM-based methods fall into two main categories: prompt engineering and model fine-tuning. Prompt engineering frameworks include decomposing test objectives into sub-questions Wang et al. (2024), offering extra static metrics to LLM in prompt Sepidband et al. (2025), and iteratively incorporating execution feedback into prompt to emulate human debugging Pizzorno & Berger (2025); Chen et al. (2024); Cheng et al. (2025). Alternatively, LLMs can be fine-tuned for the specific task of test generation Ma et al. (2025); Eom et al. (2024). In this work, we propose a reinforcement learning-based strategy that robustly integrates dynamic execution feedback with static code analysis to produce high-quality unit tests.

3 METHOD

In this section, we present CVeDRL, which has both performance and efficiency advantages at a scale of 0.6B. We begin by introducing the Unit Test Majority-Voting Framework and its associated reliability bound of test case quality. Then we introduce the base code verifier trained by syntax and functionality rewards for GRPO. To handle boundary conditions, we apply an exponential reward shaping mechanism that amplifies rewards for covering rare branches. Additionally, we integrate two static analysis metrics that provide priori code sample complexity assessments, refining the reward function with static insights.

162 3.1 UNIT TEST MAJORITY-VOTING FRAMEWORK AND CONFIDENCE BOUND
163

164 The unit test majority-voting framework adopts the well-established best-of- N decoding strat-
165 egy, wherein an LLM policy model first generates N candidate programs for a given program-
166 ming problem Q (Cobbe et al., 2021; Lightman et al., 2024a). Formally, we denote these can-
167 didates as $\{s_1, s_2, \dots, s_N\}$. To assess functional correctness, an auxiliary LLM produces M unit
168 tests for each (Q, s_i) pair, yielding test suites $\{T_1, T_2, \dots, T_M\}$, where each test suite $T_j =$
169 $\{(x_{j,1}, y_{j,1}), (x_{j,2}, y_{j,2}), \dots, (x_{j,K_j}, y_{j,K_j})\}$ contains K_j input–output pairs. Here, $x_{j,k}$ is the k -
170 th input and $y_{j,k}$ the expected output. Each candidate C_i is executed against all tests T_j , producing
171 binary outcomes

$$172 \quad p_{i,j} = \begin{cases} 1, & \text{if } s_i \text{ passes every case in } T_j, \\ 173 & 0, \quad \text{otherwise.} \end{cases}$$

174 These results form a reward vector $\mathbf{p}_i = (p_{i,1}, \dots, p_{i,M})$ for each candidate. Finally, under the
175 majority-voting criterion (Wang et al., 2023), we select the candidate that maximizes the total
176 number of passed tests:

$$177 \quad s_{\text{opt}} = \arg \max_{i \in \{1, \dots, N\}} \sum_{j=1}^M p_{i,j}.$$

180 To quantify how test-assertion reliability and branch coverage jointly influence the confidence in the
181 selected program, we derive that the assertion correctness probability p must satisfy the bound

$$183 \quad p \geq \frac{1 + \sqrt{\frac{2}{M} \ln\left(\frac{1-q}{1-q'} N\right)}}{1+c},$$

186 where q' denotes the overall probability of correctly selecting a program under the majority-voting
187 framework, q denotes the prior probability that any individual candidate is functionally correct. N
188 indicates the total number of generated programs under consideration, M signifies the number of
189 independent test suites executed for each candidate, while parameter c captures the average branch
190 coverage. By making these dependencies explicit, the bound provides practical guidelines on how
191 to trade off candidate-pool size, test-suite count, and test-generation quality to achieve a target post-
192 selection confidence level q' . This result is obtained by applying Hoeffding inequality to the bounded
193 difference of vote counts of correctness, together with a union bound over all candidate comparisons.
194 The detailed proof is in Appendix A.

195 3.2 BASE CODE VERIFIER BY REINFORCEMENT LEARNING
196

197 Previous studies Xie et al. (2025); Zeng et al. (2025) demonstrate that the design of 'format-answer'
198 reward effectively standardizes the format and ordering of model outputs, which guides the rea-
199 soning process of the model. Inspired by these studies, we propose the basic design of reinforce-
200 ment fine-tuning framework within CVeDRL that dynamic testing feedback via two complementary
201 rewards: Syntax Reward and Functionality Reward. Syntax reward enforces specific AST-derived
202 formatting rules, and Functionality Reward is based on the testing execution results. This approach
203 significantly mitigates the propensity of models to 'hack' evaluations, standardizing generated code
204 format with dynamic test execution signals.

205 **Syntax Reward.** Inspired by Guo et al. (2025), we reexamine the formatting and syntax require-
206 ments for the unit-testing task and designed a corresponding syntax-based reward. Specifically, the
207 generated tests must be strictly enclosed within a final Python code block, and traversing the AST
208 of this test code must reveal at least one class inheriting from `unittest.TestCase`. Given that
209 W_i is the i^{th} concatenation response of correct code solution s , the format reward is calculated as
210 follows:

$$211 \quad r_{\text{syn}}(W_i|s) = \begin{cases} 1.0, & \text{if syntax is correct,} \\ 212 & -1.0, \quad \text{if syntax is incorrect.} \end{cases}$$

214 **Functionality Reward.** Once the syntax is validated, a regex-based extractor retrieves the cor-
215 rectly structured unit-test snippet u_i from the response of model W_i . To tailor concepts from
prior RL-based fine-tuning methods for reducing hallucination rates specifically, we refine these

approaches by classifying actual testing execution outcomes to derive reward signals. During unit-test generation, models exhibit two primary types of hallucinations: (1) Errors, where wrong code or invalid inputs cause test execution to crash, i.e. both false positives and false negatives; and (2) Failures, where tests run successfully but incorrectly predict the output from source code s . Because failures are inherently less severe than errors, we assign distinct negative rewards proportional to each category. For tests is passed, we introduce branch coverage rate for the reward. The execution-result reward is computed as follows:

$$r_{func}(u_i, s) = \begin{cases} -2.0, & \text{if } u_i \text{ is error towards } s \\ -1.5, & \text{if } u_i \text{ is failure towards } s \\ +\text{cov}(u_i, s), & \text{if } u_i \text{ is passed} \end{cases}$$

In light of GRPO Shao et al. (2024), for question $q = q(s)$ with corresponding code solution given as the input to LLM, the model samples a group of outputs $\{o_i\}_{i=1}^G$ from the old policy $\pi_{\theta_{old}}$ then optimize the policy with the following objective iteratively:

$$\begin{aligned} \mathcal{J}(\theta, \{o_i\}_{i=1}^G) &= \mathbb{E}_{q \sim P(Q), \{o_i\} \sim \pi_{\theta_{old}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^G \min \left[r_i(\theta, \theta_{old}|q) a_{o_i}, \text{clip}(r_i(\theta, \theta_{old}|q), \varepsilon) a_{o_i} \right] \right] \\ &\quad - \beta \mathbb{E}_{q \sim P(Q), \{o_i\} \sim \pi_{\theta_{old}}(\cdot|q)} [D_{KL}(\pi_\theta \| \pi_{ref})], \\ r_i(\theta, \theta_{old}|q) &= \frac{\pi_\theta(o_i | q)}{\pi_{\theta_{old}}(o_i | q)}, \end{aligned}$$

where $\text{clip}(x, \varepsilon) := \min(\max(x, 1 - \varepsilon), 1 + \varepsilon)$, π_θ is the optimal policy. a_{o_i} is the advantage term of i^{th} output calculated by the following formula:

$$a_{o_i} = \frac{r_i - \mu_{r_i}}{\sigma_{r_i}}, r_i = \begin{cases} r_{syn}(W_i|s), & \text{if syntax is incorrect} \\ r_{syn}(W_i|s) + r_{func}(u_i, s), & \text{if syntax is correct} \end{cases}$$

where μ_{r_i} and σ_{r_i} is the mean and standard deviation.

3.3 BRANCH-DIFFICULTY-AWARE REINFORCEMENT LEARNING

The Base Code Verifier employs a linear coverage-based reward, which leads the model to favor generating happy-path test cases and to overlook many boundary conditions. We attribute this behavior to the diminishing reward at high coverage levels, resulting in insufficient exploration incentives. To encourage the model to focus more on boundary conditions and other atypical execution paths, we adopt an empirical yet efficient reward design with coverage awareness:

$$r_{cov}(u_i, s) = (e^\alpha - 1)^{-1} [\exp(\alpha \times \text{cov}(u_i, s)) - 1]$$

where $\text{cov}(u_i, s) \in [0, 1]$ is the coverage rate of unit-test case with a given code, and $\alpha > 0$ is a tunable hyperparameter controlling the tail-heaviness of the curve. r_{cov} remains nearly linear for small α but stays flat at low coverage and then rises sharply as α increases. Under this scheme, branches with low coverage (i.e. rare or non-typical paths) receive a much larger incremental reward compared to those already well covered, thereby shifting the focus of policy toward discovering boundary conditions and other special cases after the main path is mastered. By integrating $r_{cov}(u_i, s)$ into our reward signal, we empirically observe that the model devotes additional generation effort to hard-to-reach logic, resulting in more balanced and comprehensive unit-test suites.

3.4 SAMPLE-DIFFICULTY-AWARE REINFORCEMENT LEARNING

A key limitation of relying solely on dynamic feedback, such as pass/fail signals or code coverage metrics, is that these measures only become available **after** the model has already generated and attempted to execute a candidate solution. Consequently, they offer no guidance in distinguishing between “trivial failures” (e.g., minor input mismatches or off-by-one errors) and genuinely challenging code fragments that demand deeper reasoning. To address this gap, we introduce a static difficulty prior that quantifies intrinsic complexity of each solution **before** any execution takes place. Specifically, for all code solution, we compute two complementary static metrics of each code during preprocessing: Halstead Complexity(D_H) and the difficulty of Maintainability(D_M).

270 Halstead Complexity, by quantifying the variety and frequency of operators and operands, serves as
 271 a metric for the cognitive load required to parse and is widely utilized for code difficulty analysis
 272 Hariprasad et al. (2017). The formulation is $\hat{D}_H = \frac{\eta_1}{2} \times \frac{N_2}{\eta_2}$, where η_1 and η_2 denote the counts
 273 of distinct operators and operands in the source code, and N_2 is the total number of operand oc-
 274 currences. This metric captures the cognitive burden imposed by syntactic diversity and operand
 275 repetition. To mitigate the influence of extreme outliers, we collect the set of all \hat{D} values across
 276 our training corpus and determine the 95 percentile with a clipped min–max normalization \hat{D}_{95} . We
 277 then perform a clipped min–max normalization to get Halstead difficulty D_H :
 278

$$D_H = \min(\hat{D}, \hat{D}_{95})/\hat{D}_{95}.$$

281 Maintainability Index (MI) Coleman et al. (1994) further captures the anticipated effort and risk of
 282 future modifications, which is a four-metric polynomial equation for measuring how maintainable
 283 the code is. Recent work Zheng et al. (2024) on MI is applied to the study of code sample differen-
 284 tiation benchmark. The formulation is as follows: $MI = \max\{0, 100 \times (171 - 5.2 \ln V - 0.23G -$
 285 $16.2 \ln L + 50 \sin(\sqrt{2.4C})/171\}$, where V denotes the Halstead Volume, which quantifies the size
 286 and information content of the code; G is the Cyclomatic Complexity; L represents lines of code and
 287 C is the percentage of comment lines. Since higher MI values indicate easier maintenance (lower
 288 difficulty), we invert and rescale it into a unified difficulty measure of maintainability:
 289

$$D_M = \max\{0, 1 - \frac{MI}{100}\} \in [0, 1].$$

291 Low-maintainability code often deviates from the idiomatic patterns observed during pretraining,
 292 thereby undermining the transferability of learned representations and increasing hallucination rates.
 293 To emphasize the synergy between code comprehension and maintenance difficulty, we deviate a
 294 geometric mean of two metrics D_H and D_M :
 295

$$D = \sqrt{D_H \times D_M} \in [0, 1]$$

297 where indicates the co-occurrence of high values in both dimensions, i.e. only code that is both hard
 298 to comprehend and hard to maintain yields a large difficulty.
 299

300 Motivated by the imperative to distinguish genuinely challenging code fragments from trivial fail-
 301 ures prior to execution, we introduce the static difficulty D into our reward. Formally, we define the
 302 total augmented execution reward r_{func} with exponential reward shaping in coverage:
 303

$$r_{func}(u_i, s) = \begin{cases} -2.0, & \text{if } u_i \text{ is error towards } s; \\ -1.0 - (1 - D), & \text{if } u_i \text{ is failure towards } s; \\ r_{cov}(u_i, s) \cdot (1 + D), & \text{if } u_i \text{ is passed,} \end{cases}$$

306 where LLM receives larger positive feedback when the execution is passed with harder code func-
 307 tions, while failures incur softened penalties. This design ensures that the agent allocates a greater
 308 exploration effort and learning capacity to high-complexity regions of the code space, thereby im-
 309 proving overall policy robustness on complex unit-testing tasks with static difficulty awareness.
 310

311 4 EXPERIMENTS

313 4.1 EXPERIMENTAL SETUP

315 Our experiment includes two components: Validation-Coder Performance and CVeDRL Test Quality
 316 evaluation. In the evaluation of validation in coder, we employ unit-test reward signals to select
 317 among candidate code solutions produced by coder models, which reveals effectiveness of each
 318 model on the validation-coder task. In the test quality evaluation, we execute all generated unit tests
 319 and report metrics for each large model’s output, thereby providing a direct assessment of test-case
 320 quality (see in Table 2). We select a compact 0.6B parameter model for CVeDRL training primarily
 321 to ensure it can serve as a sufficiently fast and efficient code verifier.
 322

323 **Datasets.** We utilize four benchmarks to comprehensively evaluate unit-test generation: HU-
 MANEVAL+ (Liu et al., 2023b), MBPP+ (Liu et al., 2023b), LIVECODEBENCH (Jain et al., 2024),

Policy Model					
Method	Scale	Llama3 8B	Llama3 70B	GPT-3.5GPT-4o-m	
HumanEval+					
Vanilla	–	53.43	73.10	67.44	82.57
MBR-E	–	60.18	75.47	70.53	85.31
CodeT	–	65.22	76.14	73.92	85.52
Llama3.1	70B	71.95	78.41	79.88	85.48
CodeRM	8.0B	<u>72.13</u>	<u>78.66</u>	78.13	86.49
Qwen3	0.6B	55.75	73.91	69.16	83.01
Qwen3	32B	70.64	78.52	78.45	<u>86.58</u>
CVeDRL	0.6B	72.14	78.72	78.96	87.05
MBPP+					
Vanilla	–	49.17	69.28	70.15	71.32
MBR-E	–	49.98	69.75	70.49	72.12
CodeT	–	59.17	69.88	69.93	73.28
Llama3.1	70B	65.24	71.77	75.87	75.01
CodeRM	8.0B	<u>66.63</u>	72.53	<u>75.98</u>	<u>75.18</u>
Qwen3	0.6B	51.01	70.19	72.86	73.74
Qwen3	32B	65.43	<u>72.61</u>	75.71	75.04
CVeDRL	0.6B	66.79	73.60	76.21	76.93
LiveCodeBench					
Vanilla	–	11.74	25.19	20.48	34.90
MBR-E	–	12.12	25.18	20.51	34.79
CodeT	–	12.59	25.92	20.60	35.11
Llama3.1	70B	13.33	28.39	22.79	38.61
CodeRM	8.0B	<u>15.24</u>	27.81	21.81	<u>39.18</u>
Qwen3	0.6B	12.52	25.47	21.91	37.18
Qwen3	32B	15.17	27.95	<u>22.86</u>	38.51
CVeDRL	0.6B	16.75	<u>27.96</u>	23.09	39.31

Table 1: The result for code verification of CVeDRL and other baselines over three code generation benchmarks. The top two performances for each dataset and policy model are marked in **bold** and underlined.

and LEETCODE greengerong (2023). Following the methodology of CodeRM Ma et al. (2025), we select 168 function-style problems from LiveCodeBench created between January and September 2024. These tasks are known to challenge large models, making them suitable for stress-testing reward signals. To assess generation performance in broader contexts, we filter 2,360 LeetCode problems following VALTEST Taherkhani & Hemmati (2024) to exclude system-design questions, class-based interfaces, and interactive I/O. We yield 542 problems each defined by a single Python function signature with a return value. During training, we use the CODERM(Ma et al., 2025) dataset, which contains over 50,000 questions selected through a dedicated processing pipeline, and ensures that most of the chosen large model solutions are both appropriate and correct.

Metrics. We report five metrics for test quality evaluation:(1) Pass Rate (PR): the proportion of tests that execute without assertion errors. (2) Failure Rate (FR): the probability that a generated test runs successfully but the model’s expected output is incorrect. (3) Error Rate (ER): the probability that the testing doesn’t give any coverage or failure report due to errors. (4) Branch Coverage (BC): the ratio of executed branches or lines covered by the tests. (5) Assertion Number (AN): the number of assertions within the unittest class. For the evaluation of validation, we adopt the pass-of-n metric under a unit-test majority voting framework, where the final solution is selected based on the number of test cases it successfully passes.

Baseline Models. In the evaluation of Validation-Coder Performance, we select four high-accuracy policy models including GPT-3.5, GPT-4o-mini Achiam et al. (2023), LLaMA-70B Dubey et al. (2024), and LLaMA-8B Dubey et al. (2024). For reward model inference, the baselines include CodeT Chen et al. (2023), MBR-E Shi et al. (2022a), and CodeRM-8B Ma et al. (2025). We also evaluate LLaMA-70B as a reward model to compare performance against larger unaligned models, and vanilla method with random selection of code solutions. For the evaluation of test quality, we employ GPT-4o Achiam et al. (2023), GPT-3.5 Brown et al. (2020), and LLaMA3-8B itellama3-

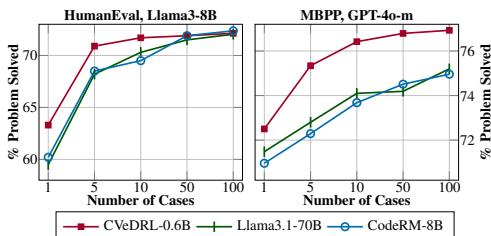
378 report as baseline models. To assess the gains provided by the CVeDRL training method, we report
 379 baseline model performance on both the direct evaluation and application evaluation tasks. We adopt
 380 Qwen3-0.6B-Base as the backbone model for training CVeDRL.
 381

382 4.2 MAIN RESULTS 383

384 **Validation-Coder Performance.** Table 1 summarizes the outcomes of our code-verification ex-
 385 periments across three benchmark datasets. Following the majority-voting scheme, each model’s
 386 candidate solutions are verified by running a fixed set of generated test cases. Compared to the pre-
 387 trained Qwen3-0.6B-base model, CVeDRL-0.6B significantly improves the best-of- n performance
 388 on all benchmarks. Against the LLaMA-70B with much larger scale, CVeDRL-0.6B yields greater
 389 gains in nearly every experiment, demonstrating that its generated unit tests more effectively dis-
 390 criminate correct solutions. For example, when using CVeDRL-0.6B as the reward model to select
 391 among GPT-4o-mini outputs on the MBPP+ dataset, the accuracy rises from 71.32% to 76.93%, an
 392 improvement of 5.61%. CVeDRL attains first-tier performance across all evaluated policy models
 393 and datasets, with most of its metrics lying within a narrow margin of the second-best baselines.
 394 These findings indicate a potential performance ceiling inherent to the policy models and demon-
 395 strate that CVeDRL, as a code verifier, can effectively select superior solutions to further elevate
 396 policy model outputs. Although larger models (e.g., 1B or 3B parameters) could potentially offer
 397 improved accuracy, the chosen 0.6B scale strikes an optimal balance between verification perfor-
 398 mance and inference efficiency.

399 **Test Quality of CVeDRL.** Combining coverage reports with static code-analysis tools, we di-
 400 rectly executed the generated tests and computed five evaluation metrics, the results of which are
 401 summarized in Table 2 for three benchmark datasets. By incorporating negative rewards for as-
 402 ssertion errors and positive, difficulty-weighted coverage rewards during training, CVeDRL-0.6B
 403 maintains an exceptionally low failure rate and high line coverage, thereby robustly improving the
 404 quality of generated unit tests. On the MBPP+ dataset, CVeDRL-0.6B significantly outperforms ev-
 405 ery compared model, with the pass rate (PR) exceeding that of the next best model GPT-4o by 16%.
 406 Notably, CVeDRL-0.6B also produces significantly fewer assertions, indicating that our method
 407 minimizes redundant test cases while sustaining coverage levels and thus further reduces testing
 408 overhead. Moreover, Table 2 shows that CVeDRL-0.6B produces the highest-quality test suites on
 409 MBPP+, and correspondingly achieves the best reward-model performance from table 1 on that task
 410 compared to all baselines. This correlation between test-generation quality and reward-model effec-
 411 tiveness suggests a tight link between coverage gain and solution selection, which we analyze in
 412 detail in Appendix A.

413 4.3 SAMPLING EFFICIENCY OF CVeDRL 414



423 Figure 2: Performance of three unit-test generators at different test-case scales, with LLaMA3-
 424 8B on HumanEval+ and GPT-4o-mini on MBPP+ as policy model separately.
 425
 426

427 Figure 2 illustrates the performance of three unit-test generators under varying test-suite scales
 428 across two policy models and datasets. Both plots demonstrate that increasing the number of sam-
 429 pled test cases generally improves reward-model efficacy, although the magnitude of improvement
 430 differs. Notably, CVeDRL-0.6B reaches its performance plateau with as few as 10 test cases in both

LLM	Scale	AMU \downarrow (GB)	AL \downarrow (s/iter)	TT \uparrow (tok/kPar/s)	AEC \downarrow (W)
CodeRM	8.0B	43.07	1.7823	0.296	293.6
Qwen3	4.0B	41.70	4.1951	0.529	294.0
Qwen3	0.6B	40.97	1.3415	6.752	276.4
CVeDRL	0.6B	40.97	0.6622	7.083	246.2

427 Table 3: Inference efficiency comparison across
 428 models. For each model, we compare four key
 429 metrics: Average Memory Utilization (AMU),
 430 Average Latency (AL), Tokens Throughput
 431 (TT), and Average Energy Consumption (AEC).
 432 All results were obtained on the MBPP+ bench-
 433 mark with 10 inference sampling per question.

432 experimental settings, indicating minimal gains when sampling up to 100 cases under a majority-
 433 voting selection criterion. When using LLaMA3-8B as the policy model on HumanEval, all three
 434 reward models perform comparably at a scale of 100, but CVeDRL-0.6B significantly outperforms
 435 the other two baselines at a scale of 5. Similarly, with GPT-4o-mini as policy model on MBPP+,
 436 CVeDRL-0.6B consistently surpasses the other reward models at every scale. This is because the
 437 unit tests generated by CVeDRL are of high quality. As derived from the confidence bound, fewer
 438 majority-voting candidates are required, which accelerates the inference process. The detailed dis-
 439 cussion is provided in the Appendix A.

440 441 4.4 INFERENCE EFFICIENCY OF CVeDRL

442 The results in Table 3 reveal clear scale-dependent trade-offs between resource utilization, latency,
 443 per-parameter throughput and energy efficiency. Relative to the SFT-trained CodeRM, CVeDRL-
 444 0.6B boosts token throughput from 0.296 to 7.083, which is an improvement of more than 20x.
 445 CVeDRL-0.6B further halves the latency from 0.66s/iter to 1.34s/iter compared with base model and
 446 marginally increases per-parameter throughput. CVeDRL-0.6B maintains competitive performance
 447 while significantly reducing computational requirements, thereby achieving a favorable trade-off
 448 between precision and latency, which aligns with the primary objective of rapidly verifying large
 449 amounts of generated code.

450 451 4.5 ABLATION STUDY

453 454 455 Model	SYN	BDA	SDA	456 Mbpp+			457 Humaneval+		
				458 PR %↑	459 BC %↑	460 AN ↓	461 PR %↑	462 BC %↑	463 AN ↓
464 CodeRM	465 –	466 –	467 –	468 44.70	469 97.11	470 7.88	471 32.83	472 96.97	473 7.15
474	475 ×	476 ×	477 ×	478 51.53	479 85.11	480 3.23	481 41.99	482 92.41	483 3.17
484 CVeDRL	485 ✓	486 ×	487 ×	488 69.47	489 89.79	490 3.75	491 68.13	492 94.83	493 3.64
496	497 ✓	498 ×	499 ✓	500 79.96	501 92.14	502 3.47	503 84.14	504 94.55	505 3.88
508	509 ✓	510 ✓	511 ×	512 71.42	513 96.75	514 3.08	515 79.15	516 97.41	517 2.75
521	522 ✓	523 ✓	524 ✓	525 83.68	526 97.37	527 3.13	528 85.94	529 97.53	530 2.41

462 Table 4: Ablation study on the MBPP+ benchmark to disentangle the contributions of the exponent-
 463 shaped exploration for syntax reward(“SYN”), branch-difficulty aware RL(“BDA”) and the static
 464 analysis reward for sample-difficulty aware RL(“SDA”).

466 We perform an ablation on CVeDRL-0.6B (Table 4) to isolate the effects of our exponent-shaped
 467 exploration for syntax reward, Branch-Difficulty aware RL and static analysis reward for Sample-
 468 Difficulty aware RL. For comparison of training methodologies, we also evaluate CodeRM-8B, a
 469 model trained on the same dataset but using supervised fine-tuning as training method. Static alone
 470 substantially improves pass rate by steering the agent toward appropriately difficult tasks, while ex-
 471 ponential shaping significantly boosts code coverage by rewarding exploration of sparsely exercised
 472 branches. Significantly, combining both yields the highest overall performance, demonstrating their
 473 complementary benefits.

474 475 476 5 CONCLUSION

477 We introduce CVeDRL, a unified reinforcement learning framework that leverages static code com-
 478 plexity metrics and test results to guide effective unit-test generation and code solution verifi-
 479 cation. Through theoretical analyses, we establish explicit relationships between test-case pass rates,
 480 branch coverage, and overall verification performance. By integrating syntax and functionality con-
 481 siderations with branch-sample difficulty-aware mechanism, CVeDRL adeptly generates concise
 482 and comprehensive test suites capable of generalizing effectively across branches and complex code
 483 solutions. Experimental verification of diverse code generation benchmarks illustrates that CVeDRL
 484 achieves top-tier performance among various policy models. Overall, CVeDRL exemplifies the po-
 485 tential of reinforcement learning in advancing the efficiency and reliability of unit test generation,
 laying a strong foundation for future research directions in code verifiers.

486 REFERENCES
487

488 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-
489 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical
490 report. *arXiv preprint arXiv:2303.08774*, 2023. URL <https://arxiv.org/abs/2303.08774>.

491 Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and
492 Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling.
493 *arXiv preprint arXiv:2407.21787*, 2024.

494 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal,
495 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and Sandhini Agarwal. Lan-
496 guage models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.

497 Alexander E. I. Brownlee, James Callan, Karine Even-Mendoza, Alina Geiger, Carol Hanna, Justyna
498 Petke, Federica Sarro, and Dominik Sobania. *Enhancing Genetic Improvement Mutations Us-
499 ing Large Language Models*, pp. 153–159. Springer Nature Switzerland, December 2023.
500 ISBN 9783031487965. doi: 10.1007/978-3-031-48796-5_13. URL http://dx.doi.org/10.1007/978-3-031-48796-5_13.

501 Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu
502 Chen. Codet: Code generation with generated tests. In *The Eleventh International Conference on
503 Learning Representations*, 2023. URL <https://openreview.net/forum?id=ktrw68Cmu9c>.

504 Yinghao Chen, Zehao Hu, Chen Zhi, Junxiao Han, Shuiguang Deng, and Jianwei Yin. Chatunitest:
505 A framework for llm-based test generation, 2024. URL <https://arxiv.org/abs/2305.04764>.

506 Xiang Cheng, Fan Sang, Yizhuo Zhai, Xiaokuan Zhang, and Taesoo Kim. Rug: Turbo Llm for Rust
507 Unit Test Generation . In *2025 IEEE/ACM 47th International Conference on Software Engineer-
508 ing (ICSE)*, pp. 2983–2995, Los Alamitos, CA, USA, May 2025. IEEE Computer Society. doi:
509 10.1109/ICSE55347.2025.00097. URL <https://doi.ieeecomputersociety.org/10.1109/ICSE55347.2025.00097>.

510 Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep
511 reinforcement learning from human preferences, 2023. URL <https://arxiv.org/abs/1706.03741>.

512 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
513 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to
514 solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL <https://arxiv.org/abs/2110.14168>.

515 Don Coleman, Dan Ash, Bruce Lowther, and Paul Oman. Using metrics to evaluate software system
516 maintainability. *Computer*, 27(8):44–49, August 1994. ISSN 0018-9162. doi: 10.1109/2.303623.
517 URL <https://doi.org/10.1109/2.303623>.

518 Heng Dong, Kefei Duan, and Chongjie Zhang. Enhancing decision-making of large language models
519 via actor-critic, 2025. URL <https://arxiv.org/abs/2506.06376>.

520 Shihan Dou, Yan Liu, Haoxiang Jia, Limao Xiong, Enyu Zhou, Wei Shen, Junjie Shan, Caishuang
521 Huang, Xiao Wang, Xiaoran Fan, Zhiheng Xi, Yuhao Zhou, Tao Ji, Rui Zheng, Qi Zhang, Xuan-
522 jing Huang, and Tao Gui. StepCoder: Improve code generation with reinforcement learning from
523 compiler feedback, 2024. URL <https://arxiv.org/abs/2402.01391>.

524 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
525 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.
526 *arXiv preprint arXiv:2407.21783*, 2024. URL <https://arxiv.org/abs/2407.21783>.

527 Jueon Eom, Seyeon Jeong, and Taekyoung Kwon. Fuzzing javascript interpreters with coverage-
528 guided reinforcement learning for llm-based mutation. In *Proceedings of the 33rd ACM SIGSOFT
529 International Symposium on Software Testing and Analysis, ISSTA 2024*, pp. 1656–1668, New
530 York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706127. doi: 10.
531 1145/3650212.3680389. URL <https://doi.org/10.1145/3650212.3680389>.

540 Juan Pablo Galeotti, Gordon Fraser, and Andrea Arcuri. Improving search-based test suite genera-
 541 tion with dynamic symbolic execution. In *IEEE International Symposium on Software Reliability
 542 Engineering (ISSRE)*, pp. 360–369. IEEE, 2013.

543 greengerong. leetcode. Hugging Face Datasets, 2023. URL <https://huggingface.co/datasets/greengerong/leetcode>.

544 Daya Guo, Dejian Yang, Huawei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
 545 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
 546 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL <https://arxiv.org/abs/2501.12948>.

547 T Hariprasad, G Vidhyagaran, K Seenu, and Chandrasegar Thirumalai. Software complexity analy-
 548 sis using halstead metrics. In *2017 International Conference on Trends in Electronics and Infor-
 549 matics (ICET)*, pp. 1109–1113, 2017. doi: 10.1109/ICOEI.2017.8300883.

550 Mark Harman and Bryan F. Jones. Search-based software engineering. *Information and Software
 551 Technology*, 43(14):833–839, 2001. doi: 10.1016/S0950-5849(01)00127-5.

552 Mark Harman, Sahar A. Mansouri, and Yuanyuan Zhang. Search-based software engineering:
 553 Trends, techniques and applications. *ACM Computing Surveys*, 45(1), December 2012. doi:
 554 10.1145/2362456.2362458.

555 Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Zilin Zhu, Weixun Wang, Songlin Jiang, Hao-
 556 ran Wang, Hao Chen, Bin Chen, Weikai Fang, Xianyu, Yu Cao, Haotian Xu, and Yiming
 557 Liu. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework, 2025. URL
 558 <https://arxiv.org/abs/2405.11143>.

559 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec
 560 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint
 561 arXiv:2412.16720*, 2024.

562 Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando
 563 Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free
 564 evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024. URL
 565 <https://arxiv.org/abs/2403.07974>.

566 Zongze Jiang, Ming Wen, Jialun Cao, Xuanhua Shi, and Hai Jin. Towards understanding
 567 the effectiveness of large language models on directed test input generation. In *Proceed-
 568 ings of the 39th IEEE/ACM International Conference on Automated Software Engineering,
 569 ASE ’24*, pp. 1408–1420, New York, NY, USA, 2024. Association for Computing Machinery.
 570 ISBN 9798400712487. doi: 10.1145/3691620.3695513. URL <https://doi.org/10.1145/3691620.3695513>.

571 Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton
 572 Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling
 573 reinforcement learning from human feedback with ai feedback, 2024. URL <https://arxiv.org/abs/2309.00267>.

574 Caroline Lemieux, Jeevana Priya Inala, Shuvendu K. Lahiri, and Siddhartha Sen. Codamosa: Es-
 575 caping coverage plateaus in test generation with pre-trained large language models. In *Proceed-
 576 ings of the 45th International Conference on Software Engineering, ICSE ’23*, pp. 919–931.
 577 IEEE Press, 2023. ISBN 9781665457019. doi: 10.1109/ICSE48619.2023.00085. URL <https://doi.org/10.1109/ICSE48619.2023.00085>.

578 Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan
 579 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The
 580 Twelfth International Conference on Learning Representations*, 2024a. URL <https://openreview.net/forum?id=v8L0pN6EOi>.

581 Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan
 582 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The
 583 Twelfth International Conference on Learning Representations*, 2024b. URL <https://openreview.net/forum?id=v8L0pN6EOi>.

594 Jiate Liu, Yiqin Zhu, Kaiwen Xiao, Qiang Fu, Xiao Han, Wei Yang, and Deheng Ye. Rltf: Rein-
 595forcement learning from unit test feedback, 2023a. URL <https://arxiv.org/abs/2307.04349>.
 596

597 Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and LINGMING ZHANG. Is your code generated
 598 by chatGPT really correct? rigorous evaluation of large language models for code generation.
 599 In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL <https://openreview.net/forum?id=1qvxf610Cu7>.
 600

601 Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu.
 602 Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*, 2025.
 603

604 Zeyao Ma, Xiaokang Zhang, Jing Zhang, Jifan Yu, Sijia Luo, and Jie Tang. Dynamic scaling of unit
 605 tests for code reward modeling, 2025. URL <https://arxiv.org/abs/2501.01054>.
 606

607 Phil McMinn. Search-based software test data generation: a survey. *Software Testing, Verification
 608 and Reliability*, 14(2):105–156, 2004. doi: 10.1002/stvr.295.

609 Juan Altmayer Pizzorno and Emery D. Berger. Coverup: Effective high coverage test generation for
 610 python, 2025. URL <https://arxiv.org/abs/2403.16218>.
 611

612 Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and
 613 Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model,
 614 2024. URL <https://arxiv.org/abs/2305.18290>.
 615

616 Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa,
 617 Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free
 618 rlhf, 2024. URL <https://arxiv.org/abs/2405.20304>.
 619

620 Melika Sepidband, Hamed Taherkhani, Song Wang, and Hadi Hemmati. Enhancing llm-based code
 621 generation with complexity metrics: A feedback-driven approach, 2025. URL <https://arxiv.org/abs/2505.23953>.
 622

623 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
 624 Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathe-
 625 matical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.
 626

627 Freda Shi, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer, and Sida I. Wang. Natural
 628 language to code translation with execution. In Yoav Goldberg, Zornitsa Kozareva, and
 629 Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Lan-
 630 guage Processing*, pp. 3533–3546, Abu Dhabi, United Arab Emirates, December 2022a. As-
 631 sociation for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.231. URL <https://aclanthology.org/2022.emnlp-main.231>.
 632

633 Freda Shi, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer, and Sida I. Wang. Natural
 634 language to code translation with execution. In Yoav Goldberg, Zornitsa Kozareva, and
 635 Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Lan-
 636 guage Processing*, pp. 3533–3546, Abu Dhabi, United Arab Emirates, December 2022b. As-
 637 sociation for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.231. URL <https://aclanthology.org/2022.emnlp-main.231/>.
 638

639 Hamed Taherkhani and Hadi Hemmati. Valtest: Automated validation of language model generated
 640 test cases, 2024. URL <https://arxiv.org/abs/2411.08254>.
 641

642 Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu,
 643 Guoyin Wang, and Eduard Hovy. Reinforcement learning enhanced llms: A survey, 2025a. URL
 644 <https://arxiv.org/abs/2412.10400>.
 645

646 Sizhe Wang, Yongqi Tong, Hengyuan Zhang, Dawei Li, Xin Zhang, and Tianlong Chen. Bpo:
 647 Towards balanced preference optimization between knowledge breadth and depth in alignment,
 2025b. URL <https://arxiv.org/abs/2411.10914>.

648 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha
 649 Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language
 650 models. In *The Eleventh International Conference on Learning Representations*, 2023. URL
 651 <https://openreview.net/forum?id=1PL1NIMMrw>.

652 Yinjie Wang, Ling Yang, Ye Tian, Ke Shen, and Mengdi Wang. Co-evolving ILM coder and unit
 653 tester via reinforcement learning, 2025c. URL <https://arxiv.org/abs/2506.03136>.

654 Zejun Wang, Kaibo Liu, Ge Li, and Zhi Jin. Hits: High-coverage ILM-based unit test generation via
 655 method slicing, 2024. URL <https://arxiv.org/abs/2408.11324>.

656 Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu,
 657 Zhirong Wu, and Chong Luo. Logic-rl: Unleashing ILM reasoning with rule-based reinforcement
 658 learning, 2025. URL <https://arxiv.org/abs/2502.14768>.

659 Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-
 660 zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025.
 661 URL <https://arxiv.org/abs/2503.18892>.

662 Jian Zhao, Runze Liu, Kaiyan Zhang, Zhimu Zhou, Junqi Gao, Dong Li, Jiafei Lyu, Zhouyi Qian,
 663 Biqing Qi, Xiu Li, et al. Genprm: Scaling test-time compute of process reward models via gener-
 664 ative reasoning. *arXiv preprint arXiv:2504.00891*, 2025.

665 Jiasheng Zheng, Boxi Cao, Zhengzhao Ma, Ruotong Pan, Hongyu Lin, Yaojie Lu, Xianpei Han,
 666 and Le Sun. Beyond correctness: Benchmarking multi-dimensional code generation for large
 667 language models, 2024. URL <https://arxiv.org/abs/2407.11470>.

672 A PASS-COVERAGE TRADE-OFF IN AUTOMATED CODE VERIFICATION

673 A.1 SINGLE-TEST ASSUMPTIONS

674 In this subsection, we derive a concise relationship between the probability that a candidate program
 675 passes a test suite and the average branch coverage of the suite. We begin by introducing the notation
 676 for the single-test-case scenario, and then generalize our findings to the multi-test context via the
 677 mean coverage.

678 Let q be the prior probability that a candidate program is *correct* and p is the probability that a single
 679 *generated test* returns PASS on correct code, i.e., that the test's own assertion is valid. r_{ij} is the
 680 probability that the test suite covers the j -th defective path (branch) of candidate i , where n_i denotes
 681 the number of defective paths in candidate i . c means branch coverage, typically approximated by
 682 $c = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^{n_i} r_{ij}$; $0 \leq c \leq 1$. Then we come up with three assumptions as follows:

683 **Model Assumptions.** We employ the following simplifying assumptions:

- 684 (a) **Correct code passes all exercised paths.** That is, if the test exercises any path of correct
 685 code, it will *pass* with probability p , and with probability $1 - p$ it fails due solely to a faulty
 686 assertion (a *false negative*).
- 687 (b) **Erroneous code has at least one faulty path.** If the test exercises one of these faulty paths
 688 and its assertion is valid, the error is revealed (the test fails as intended); otherwise the test
 689 yields a *false positive*.
- 690 (c) **Suite-level passing criterion.** A program is declared to *pass* the entire test suite and thus
 691 be emitted as a candidate if and only if it passes every individual test (the simplified
 692 `pass_of_n` rule).

693 **Single-Test Pass Probabilities.** Under these assumptions, let C be the event “program is correct”.
 694 Then for a single test we have

$$695 P(\text{pass} \mid C) = p, \quad P(\text{pass} \mid \bar{C}) = 1 - p r_i,$$

702 where

$$703 \quad r_i = \sum_{j=1}^{n_i} r_{ij}$$

706 is the total probability that the test covers any one of the defective paths in candidate i . In practice,
707 one often replaces R_i by the average coverage s , yielding the approximation

$$708 \quad P(\text{pass} \mid \bar{C}) \approx 1 - p c.$$

710 **Posterior Probability of Correctness.** Given that a candidate program has *passed* the test, we
711 apply Bayes' theorem to update our belief:

$$713 \quad P(C \mid \text{pass}) = \frac{P(\text{pass} \mid C) P(C)}{P(\text{pass} \mid C) P(C) + P(\text{pass} \mid \bar{C}) P(\bar{C})}$$

$$714 \quad = \frac{q p}{q p + (1 - q)(1 - p r_i)}.$$

718 **Threshold for Improved Posterior.** We are particularly interested in the condition under which
719 the posterior probability of correctness exceeds the prior:

$$721 \quad P(C \mid \text{pass}) > q \implies p > \frac{1}{1 + r_i}.$$

723 Substituting the mean coverage approximation $R_i \approx s$ yields the practical threshold

$$725 \quad p > \frac{1}{1 + c}.$$

727 In other words, provided that the single-test reliability p exceeds the reciprocal of one plus the
728 average coverage, passing the test suite will *increase* our confidence that the program is indeed
729 correct. This completes the derivation of the basic formula relating pass probability and coverage
730 under our simplified single-test assumptions. Extensions to multiple independent tests follow by
731 replacing p and r_i with their compounded quantities across the suite.

733 A.2 MAJORITY VOTING FRAMEWORK

735 In this section, we extend our analysis to a *majority-voting* scheme. Let N be the number of candidate
736 programs (*best-of- N*) and M be the number of independently generated test suites per candidate. K denotes the number of I/O pairs in each suite, i.e. the number of assumptions across suites.

738 **Per-Suite Pass Probability.** When each suite contains K I/O pairs, all assertions must pass for
739 the suite to be considered successful. Hence

$$741 \quad \alpha_c = p^K \quad \text{for a correct program,}$$

$$742 \quad \alpha_w = (1 - p c)^K \quad \text{for an incorrect program.}$$

744 where $1 - p c$ is the probability that either the defect path is not covered or the assertion fails to
745 detect the error. We simplify our model in $K = 1$.

747 **Binomial Model for Passing Suites.** For any fixed candidate, the number of suites that pass follows a binomial distribution:

$$749 \quad p \sim \text{Binomial}(M, \alpha)$$

750 where

$$751 \quad \alpha = \begin{cases} \alpha_c, & \text{if the program is correct,} \\ 752 \quad \alpha_w, & \text{if the program is incorrect.} \end{cases}$$

754 Under majority voting, we select the candidate with

$$755 \quad s_{\text{opt}} = \arg \max_{i=1, \dots, N} p_i.$$

756 **Reliability via Concentration Bounds.** To ensure that the probability of selecting a wrong program
 757 is at most δ , we require

$$758 \quad P(s_{\text{opt}} \text{ is correct}) \geq 1 - \delta.$$

759 We now show how this follows from a Hoeffding-type concentration inequality applied to the difference
 760 in suite-pass counts between the true correct program and any incorrect one.
 761

762 For each test index $t \in \{1, \dots, M\}$ and each wrong candidate $j \in \{1, \dots, WN\}$ where the wrong
 763 proportion of N code solutions is W , define the indicator variables

$$764 \quad x_t = \mathbf{1}_{\text{test } t \text{ passes on the correct solution}},$$

$$765 \quad y_{j,t} = \mathbf{1}_{\text{test } t \text{ passes on wrong solution } j}.$$

767 By construction these are independent and identically distributed (i.i.d.), with

$$768 \quad P(x_t = 1) = \alpha_c, \quad P(y_{j,t} = 1) = \alpha_w,$$

770 where α_c and α_w are the per-suite pass probabilities for correct and incorrect code, respectively.
 771

772 Let

$$773 \quad g_c = \sum_{t=1}^M x_t, \quad g_j = \sum_{t=1}^M y_{j,t},$$

$$776 \quad d_j = g_c - g_j = \sum_{t=1}^M (x_t - y_{j,t}) = \sum_{t=1}^M z_{j,t}$$

779 where each summand $z_{j,t} = x_t - y_{j,t}$ satisfies $z_{j,t} \in [-1, 1]$. The expectation of the gap is
 780

$$781 \quad \mathbb{E}[d_j] = \sum_{t=1}^M \mathbb{E}[z_{j,t}] = M(\alpha_c - \alpha_w) = M\Delta.$$

784 By the one-sided Hoeffding bound for bounded independent random variables,
 785

$$786 \quad P(d_j \leq 0) = P(d_j - \mathbb{E}[d_j] \leq -M\Delta) \\ 787 \quad \leq \exp\left(-\frac{2(M\Delta)^2}{\sum_{t=1}^M (1 - (-1))^2}\right) \\ 788 \quad = \exp\left(-\frac{M\Delta^2}{2}\right) =: \beta.$$

791 Hence for each wrong candidate j , the probability that it ties or outperforms the correct program is
 792 at most β .

793 Then we calculate the union bound over all wrong candidates. Let \mathcal{E} be the event that *any* of the
 794 WN wrong programs achieves $s_j \geq s_c$, which indicates s_{opt} is incorrect. By the union bound,
 795

$$796 \quad P(\mathcal{E}) \leq WN\beta = WN \exp\left(-\frac{M\Delta^2}{2}\right).$$

798 Imposing the reliability target $P(\mathcal{E}) \leq \delta$ gives

$$800 \quad WN \exp\left(-\frac{M\Delta^2}{2}\right) \leq \delta \implies \Delta \geq \sqrt{\frac{2 \ln \frac{WN}{\delta}}{M}}.$$

802 Recalling that $\Delta = \alpha_c - \alpha_w$ and $P(s_{\text{opt}} \text{ is correct}) \geq 1 - \delta$, we obtain the required ‘safety margin’
 803

$$804 \quad \alpha_c - \alpha_w \geq \sqrt{\frac{\ln \frac{WN}{\delta}}{2M}}$$

808 Operational bound for $K = 1$. In the single-case suite model ($K = 1$), we have $\alpha_c = p$ and
 809 $\alpha_w = 1 - p$, hence

$$810 \quad \Delta = p - (1 - p)c = p(1 + c) - 1.$$

Table 5: Estimated minimal number of independent test suites per candidate M (rounded) for $N = 100$ with branch coverage c of CVeDRL in Table 1, varying assertion reliability p . As an illustrative example, we adopt CVeDRL as the verifier and GPT-4o-m as the code generator, which is demonstrated in Table 2.

Dataset	$p = 0.70$	$p = 0.80$	$p = 0.85$
HumanEval+ ($q = 0.8257 \rightarrow q' = 0.8705, c = 0.96$)	71	30	21
MBPP+ ($q = 0.7132 \rightarrow q' = 0.7693, c = 0.97$)	67	29	20

Then we have

$$p(1 + c) - 1 \geq \sqrt{\frac{\ln \frac{WN}{\delta}}{2M}}.$$

Substitution yields the practical requirement

$$p \geq \frac{1 + \sqrt{\frac{2}{M} \ln(\frac{WN}{\delta})}}{1 + c}.$$

Finally, if we denote by $q' = 1 - \delta$ the overall probability of correct selection with majority voting framework and replace the fraction of wrong solutions W by its expectation $E[W] = 1 - q$, the bound becomes

$$p \geq \frac{1 + \sqrt{\frac{2}{M} \ln(\frac{1-q}{1-q'} N)}}{1 + c} \quad (*)$$

Inference setting with majority-voting framework and confidence bound. As an example, we show how the confidence bound quantifies the trade-off between test quality and the reduction in required test suites M per candidate when selecting among multiple program proposals. The numerical entries in Table 5 (q, q', c, p) is obtained from the experimental tables, where the prior q is the baseline (“Vanilla”) correctness rate measured for the policy column, and the posterior target q' is the improved correctness rate reported under CVeDRL (the “CVeDRL” row) from Table 1. The coverage parameter c was taken from the verifier branch coverage statistics (BC) in Table 2 and used as a proxy for average branch coverage (hence $c \approx 0.96$ for HumanEval+ and $c \approx 0.97$ for MBPP+). The per-assertion reliability p was treated as the variable in the sensitivity grid reported in Table 5. The operational bound (*) indicates that

$$M \approx \frac{2 \ln(\frac{1-q}{1-q'} N)}{((1 + c)p - 1)^2}.$$

Improving the verifier (as CVeDRL does) reduces the minimal required number of independent suites M through two channels: (i) raising the effective assertion reliability p (signalled by higher PR and lower ER/FR) and (ii) increasing measured coverage c (higher BC). Both actions enlarge the denominator $((1 + c)p - 1)^2$ and thus shrink M (the dependence is approximately quadratic in the effective margin). However, the inequality also highlights a countervailing effect: specifying a higher target post-selection confidence q' increases the logarithmic numerator $\ln(\frac{1-q}{1-q'} N)$, and therefore raises the required M . Fundamentally, we set $N = M = 100$ during inference time. In this case, the number of samples M required for CVeDRL to attain the target correctness rate q' is substantially less than 100. Notably, CVeDRL nearly reaches the desired performance at $M \approx 10$ as illustrated in Figure 2, which clearly surpasses the capability of other code verifiers. As shown in Table 5, CVeDRL achieves a pass rate close to 85%, with the required number of majority-voting candidates being around 20. This indicates that the target performance can be reached when $M \approx 20$, which is more than three times fewer samples than those required by a verifier with a unit test accuracy of $p = 0.70$ for both dataset.

A.3 CONCLUSION

In summary, our analysis reveals that the interplay between test-assertion reliability p and average branch coverage s fundamentally determines the posterior confidence in a candidate program’s correctness. A single test suffices to improve confidence only when $p > 1/(1 + c)$. With majority

voting framework across M independent suites and N candidates, the pass rate of unit-test generated by LLMs requires a stronger condition (\star). These bounds quantify explicit trade-offs. More test suites or higher coverage can compensate for imperfect assertions, whereas larger candidate pools or stricter error tolerances demand more reliable tests. This provides principled guidelines for designing automated testing pipelines that balance resource expenditure against desired selection reliability. Specifically, the analysis is utilized for choosing the crucial hyperparameter M which is relative to acceleration.

B ADDITIONAL DISCUSSIONS

B.1 IMPACT OF ANCILLARY LLM COMPONENTS

To investigate how the choice of base model influences the effectiveness of the CVeDRL training method, we include Qwen2.5-Coder-0.5B as a benchmark against Qwen3. Additionally, we perform an ablation study by disabling Qwen3’s chain-of-thought capability using the /no_think tag. Finally, motivated by prior findings that large LLMs struggle with static code analysis, we augment the prompts with conditional branch information to assess whether this auxiliary context yields further performance gains.

The ablation study in Table 6 indicates that prompting primarily guides the model toward correct solutions with fewer attempts, while chain-of-thought (CoT) encourages broader exploration at the expense of efficiency. In the smaller Qwen2.5 model, prompts reduce the average number of trials but slightly lower overall pass rate, whereas disabling prompts yields higher pass rate with more attempts. Although Qwen2.5 shows reduced pass rate and coverage relative to Qwen3, it requires fewer generated assertions and nonetheless significantly outperforms GPT-4o, highlighting the superiority of the CVeDRL training methodology. The ablation results show that introducing chain-of-thought (CoT) consistently harms overall performance, lowering both pass rate and coverage. Moreover, adding conditional prompts has no appreciable effect on branch coverage but nonetheless reduces the model’s success rate. These findings indicate that while CoT and extra prompt constraints aim to guide reasoning, they in fact impede efficiency without delivering coverage benefits.

Table 6: Ablation of Prompt and CoT on Validation-Coder Performance

Base	Prompt	CoT	PR%	BC%	AN
Qwen2.5	✓	\	71.68	96.79	2.09
Qwen2.5	✗	\	75.43	96.86	<u>2.53</u>
Qwen3	✗	✗	82.01	<u>97.14</u>	2.91
Qwen3	✗	✓	<u>80.37</u>	96.15	3.47
Qwen3	✓	✗	79.98	97.21	3.13
Qwen3	✓	✓	79.71	96.54	3.64
GPT-4o	\	\	66.13	96.91	6.12

B.2 LIMITATION

Partial code support. In real-world code generation tasks such as code completion, the generated code is often partial and thus cannot be directly validated by unit tests. Although the sample-branch syntax/functionality reward we propose can be adapted to handle partial programs and library-level (test-suite) training once a suitable development environment and dependencies are provisioned, this capability has not yet been integrated into our pipeline. Future work will close this gap by incorporating partial-code validation into the pipeline and extending the approach to support additional programming languages and broader library ecosystems.

Unit test adaptability. While our experiments show that the current unit tests generated by CVeDRL are effective at filtering incorrect code, the verifier itself cannot distinguish the intrinsic cor-

918
919
920
921
922
rectness of the code solution. Future work should incorporate code mutation or auxiliary principles
to alleviate this limitation.

921 C TRAINING CONFIGURATION 922

923
924
925
926
927
928
929
930
931
Framework and Algorithm. We fine-tune the Qwen3-0.6B and Qwen2.5-Coder-0.5B checkpoint
932 with the `ver1` RL library using *Group Relative Policy Optimization* (GRPO). `ver1` is an open-
933 source reinforcement-learning framework designed for post-training fine-tuning of large language
934 models, providing a hybrid single- and multi-controller programming model for scalable PPO and
935 GRPO workflows. It features modular APIs that decouple computation and data dependencies and
936 integrates seamlessly with PyTorch FSDP, Megatron-LM, vLLM, and other LLM infrastructures
937 for efficient, production-ready deployments. GRPO eliminates the separate value network and up-
938 dates the policy by comparing each sampled trajectory to the within-group reward baseline, thereby
939 reducing both memory footprint and wall-clock cost.

940
941
942
Dataset. All prompts are taken from the publicly-available CodeRM-UnitTest corpus, which pro-
943 vides 17,600 training items and 59,600 held-out test items. The CodeRM-UnitTest dataset is a cu-
944 rated collection of over 77 000 synthetic Python unit tests, derived from CodeFeedback-Filtered-
945 Instruction and TACO, and provided in Parquet format for training test-guided code-reward models.
946 It serves as the primary training and evaluation corpus for lightweight unit-test generator models like
947 CodeRM-8B, enabling rigorous performance benchmarks under real execution feedback. Because
948 every roll-out requires real execution of unit tests, we randomly subsample 3,000 test cases for val-
949 idation to keep evaluation under two hours per checkpoint. Experiments ran on 2 × NVIDIA A100
950 40 GB GPUs (FP16) with total training time of roughly 48 hours. Table 7 summarises the hyper-
951 parameters that most influence optimization and compute. All other settings follow the official `ver1`
952 GRPO recipe.

944 Parameter	945 Value
946 Learning rate	1×10^{-6}
947 Global prompt batch size	32
948 Rollouts per prompt	2
949 Max prompt / response length	6 150 / 2 048 tokens
950 Mini-batch / micro-batch size	16 / 8
951 KL loss coefficient	0.001 (low-variance)
952 Entropy coefficient	0 (disabled)
953 GRPO clip ratio	0.2 (default)
954 Total epochs	1 000
955 Gradient checkpointing	enabled
FSDP offload	disabled

956
957 Table 7: Salient hyper-parameters used in GRPO fine-tuning.
958

959 D EXPERIMENTAL DETAILS 960

961
962
963
Policy models. We evaluate four instruction-tuned policy models of different capacity and
964 provider type: *Llama3-8B-Instruct*, *Llama3-70B-Instruct*, *GPT-3.5-turbo*, and *GPT-4o-mini*. Each
965 model produces at most 100 candidate code solutions per prompt. Decoding and verification are run
966 on 4 × NVIDIA A100-40 GB GPUs.

967
Baselines. We consider three verification-oriented baselines that exploit unit-test feedback to dis-
968 criminate among candidate programs.

- 969 • Vanilla : the top-1 sample of the policy model without any reranking.
- 970 • MBR-E : minimum-Bayes-risk decoding that ranks candidates by the empirical risk com-
971 puted from the execution outcomes of LLM-generated test cases (we use the “hard-loss”
variant).

972
973
974

- CodeT : dual-execution agreement that measures both (i) consistency between candidate
programs and their generated tests and (ii) cross-candidate concordance.

975 Besides the three test-driven methods discussed above, we measure performance against four
976 capacity-oriented baselines: an 8B reward model CodeRM-8B for score-based reranking, a strong
977 70B code LLM Llama3-70B-Instruct, the supervised-fine-tuned backbone Qwen3-0.6B (Base), and
978 our model CVeDRL-0.6B.

979
980
981
982
983
984
985
986
987

- CodeRM-8B : an 8B-parameter reward model that assigns a scalar quality score to each candidate; we select the highest-scoring solution.
- Llama3.1-70B : a strong open-source coder whose single best sample is taken as a stand-alone baseline.
- Qwen3-0.6B (Base) : the supervised-fine-tuned version of our backbone model, without reinforcement learning.
- CVeDRL-0.6B : our method, trained with GRPO on weighted data (see Appendix C).

988 **Datasets.** To obtain a representative view of unit-test generation, we aggregate four publicly
989 available benchmarks: HumanEval+ and MBPP+, LiveCodeBench from Jan to Sep 2024, and the
990 algorithm-oriented subset of LeetCode. Together, they cover both synthetic interview-style problems
991 and organically authored, in-the-wild code, furnishing a diverse test bed for large-language-model
992 (LLM) evaluation.

993
994
995
996
997
998
999
1000

- LivecodeBench : Following the CodeRM protocol, we retain the 168 function-style tasks released between January – September 2024, because these newer problems have empirically proved the most challenging for current LLMs.
- LeetCode : Starting from 2,360 publicly accessible problems, we apply the VALTEST filtering rules to discard system-design, class-interface, and interactive I/O questions. The remaining 542 tasks each expose one Python function signature with a deterministic return value, enabling uniform test-harness construction.

1001 For every benchmark we normalise signatures, strip extraneous boilerplate, and compile tests into
1002 a unified execution harness so that success rate, failure rate, branch coverage, and Pass@N can be
1003 measured consistently across datasets.

1004
1005 **Experimental Rationale.** HUMANEVAL+ and MBPP+ are retained in both the CVeDRL Test
1006 Quality and Validation-Coder Performance studies in main result because their moderate task counts,
1007 authoritative reference solutions, and fine-grained coverage tooling permit reliable intrinsic scoring
1008 while keeping the compute requirements of iterative Pass@N evaluation tractable. The filtered
1009 LEETCODE subset is confined to the Test Quality analysis. Its larger problem set and costly edge-
1010 case generators provide valuable stress-testing for success, error, and coverage metrics, but render
1011 exhaustive validation-coder search computationally infeasible. Conversely, LIVECODEBENCH ap-
1012 pears only in the Validation-Coder study. Its developer-written tests supply a strong external oracle
1013 for synthesis evaluation, yet their heterogeneity prevents fair aggregation with coverage-based qual-
1014 ity metrics.

1015 Due to the constraints of the double-blind review policy, the model weights are not publicly released
1016 at this stage.

1017
1018
1019
1020
1021
1022
1023
1024
1025

```

1026
1027
1028
1029
1030
1031
1032 [Original Prompt](Input)
1033 Below is a question and it's corresponding code answer. Please write test cases to check the correctness of the code answer. You need to
1034 use the unittest library in Python and create a test class for testing.
1035
1036 ### question
1037 Subodhians are taught to add multi-digit numbers from right-to-left one digit at a time. Many find the "carry" operation - in which 1 is
1038 carried from one digit position to be added to the next - to be a significant challenge. Your job is to count the number of carry operations
1039 for each of addition problem so that educators may assess their difficulty.\n\nFor the input first line contains n number of records which is
1040 less then 1000. And then each line contains two numbers with spacing between them.
1041
1042 ### code solution
1043 def count_carry_operations(a: str, b: str) -> int:
1044     # Ensure both strings are of equal length by padding with leading zeros
1045     max_len = max(len(a), len(b))
1046     a = a.zfill(max_len)
1047     b = b.zfill(max_len)
1048
1049     carry_count = 0
1050     carry = 0
1051
1052     # Iterate from the last digit to the first
1053     for i in range(max_len - 1, -1, -1):
1054         digit_sum = int(a[i]) + int(b[i]) + carry
1055         if digit_sum >= 10:
1056             carry = 1
1057             carry_count += 1
1058         else:
1059             carry = 0
1060
1061     return carry_count
1062
1063
1064 [Unit Test Output](Model Response)
1065 Here is the test case for the `count_carry_operations` function:
1066
1067 ``python
1068 import unittest
1069
1070 class TestCountCarryOperations(unittest.TestCase):
1071     def test_example1(self):
1072         self.assertEqual(count_carry_operations("123", "456"), 0)
1073
1074     def test_example2(self):
1075         self.assertEqual(count_carry_operations("555", "555"), 3)
1076
1077     def test_example3(self):
1078         self.assertEqual(count_carry_operations("123", "594"), 1)
1079
1080     if __name__ == '__main__':
1081         unittest.main()
1082
1083 This test case checks the function with the provided sample
1084 inputs and outputs.<|endoftext|>
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179

```

Test Coverage

```

-----
... [100%]
----- coverage: platform linux, python 3.10.16-final-0 -----
Name Stmts Miss Cover Missing
-----
src.py 13 0 100%
-----
TOTAL 13 0 100%
3 passed in 0.02s

{'coverage': '100%', 'uncovered_lines': {'src.py': ''}, 'failure': False,
'error': False, 'assertion_info': {'assertEqual': 3}}
{'coverage': 1.0, 'assertion_info': {'assertEqual': 3}}


-----
Answer score: 1.0
Format score: 1.0
Total score: 2.0
-----
```

Figure 3: A case of the training pipeline for CVeDRL.