
SLMFIX: LEVERAGING SMALL LANGUAGE MODELS FOR ERROR FIXING WITH REINFORCEMENT LEARNING

David Jiahao Fu ^{*1} Aryan Gupta ^{*1} Aaron Councilman ¹ David Grove ² Yu-Xiong Wang ¹ Vikram Adve ¹

ABSTRACT

Recent advancements in large language models (LLMs) have shown very impressive capabilities in code generation across many programming languages. However, even state-of-the-art LLMs generate programs that contain syntactic errors and fail to complete the given tasks, especially for low-resource programming languages (LRPLs). In addition, high training cost makes finetuning LLMs unaffordable with constrained computational resources, further undermining the effectiveness of LLMs for code generation. In this work, we propose **SLMFix**, a novel code generation pipeline that leverages a small language model (SLM) finetuned using reinforcement learning (RL) techniques to fix syntactic errors in LLM-generated programs to improve the quality of LLM-generated programs for domain-specific languages (DSLs). In specific, we applied RL on the SLM for the program repair task using a reward calculated using both a static validator and a static semantic similarity metric. Our experimental results demonstrate the effectiveness and generalizability of our approach across multiple DSLs, achieving more than 95% pass rate on the static validator. Notably, SLMFix brings substantial improvement to the base model and outperforms supervised finetuning approach even for 7B models on a LRPL, showing the potential of our approach as an alternative to traditional finetuning approaches.

1 INTRODUCTION

Program synthesis, a process that automatically generates a program from some (hopefully higher-level) description, has attracted the attention of many researchers since almost the beginning of the modern study of computer science. In the programming languages community, this refers to the automated generation of programs from a formal specification, such as a type signature, test cases, or logical specification; and generally uses complex search algorithms to generate the program. The goal in this setting is to construct a program that provably satisfies the given specification. In parallel, in the AI community, program synthesis refers to the automated generation of programs from high-level, informal specifications, generally a natural language description, and using an AI agent to generate the program. The AI-based approach requires the agent to understand both the syntax of the language and the task description in depth and perform multi-step reasoning to create a reasonable program, making it a very challenging task for AI agents.

As transformer-based large language models (LLMs) have shown significant performance improvement in reasoning-intensive natural language tasks, researchers have explored the possibility of leveraging LLMs for code generation and

recent state-of-the-art models including GPT-5 (OpenAI, 2025) and Gemini (Gemini Team, 2025) have achieved performance comparable to top-tier human programmers (Li et al., 2022; OpenAI et al., 2025), while agentic systems like GPT Codex (Chen et al., 2021) and Claude (Anthropic, 2024) have seen wide adoption in industry.

Despite their success in code generation tasks across many programming languages, the current training and inference pipeline, that leverages supervised finetuning techniques for direct code generation, still faces several critical challenges. First, unlike a typical human programming approach, LLMs directly generate the output in the target language based on the input query without verifying the syntactic or functional correctness of the output. This approach has no safeguards against incorrect code, where even state-of-the-art (SOTA) LLMs are prone to syntactic errors, semantic mistakes, and possible model hallucinations (Song et al., 2023; Zhang et al., 2025b). Second, existing training processes require a large training dataset to fully understand the syntax of the target language. However, for low-resource programming languages (LRPLs), such datasets are often unavailable and difficult to construct, due to the scarcity and fragmentation of high-quality training data for the target language and limited community usage. As a result, LLMs tend to show inferior performance on code generation tasks involving LRPLs (Orlanski et al., 2023). Finally, finetuning on the code generation task for the specific target language is gen-

^{*}Equal contribution ¹University of Illinois Urbana-Champaign

²IBM Research.

erally required for LLMs to achieve best performance, but training is a time-consuming process with high hardware requirements.

Prior works (Zhang et al., 2023; Gong et al., 2022; Zhang et al., 2025a) have attempted to tackle these problems by incorporating a compiler and a program fixer into the inference pipeline and designing more advanced training methods, but they remain largely ineffective in mitigating syntactic errors in generated programs due to the employment of LLMs that are training-free or trained with SFT, while reinforcement learning (RL) approaches (Le et al., 2022; Liu et al., 2023; Jha et al., 2024) can generally achieve better performance but also bring higher training costs.

Researchers have proposed various techniques to address the challenge of high training cost while maintaining the effectiveness of training, but few have explored using small language models (SLMs) trained by reinforcement learning. Recent studies (Kusama et al., 2025; Sheng & Xu, 2025) have showed the potential of small language models (SLMs) for code-related tasks, and the task of automated program repair could be very suitable for SLMs because the model only needs to be able to make corrections, not generate the code from scratch.

In this paper, we are particularly interested in improving the quality of LLM-generated programs for domain-specific languages (DSLs). While general-purpose programming languages have vast datasets and well-established tooling, DSLs are often low-resource since they lack such large-scale resources, which makes reliable code generation particularly challenging. We propose a novel code generation pipeline that can be applied to any existing LLMs without finetuning on the LLM, and performs well regardless of the size of the training data available for the target language. We achieve this by incorporating an SLM to fix statically detected errors, such as syntax and type errors, in LLM-generated programs. With extensive experiments, we found that SLMs are able to perform very well on this error-fixing task; we show that a 500M parameter model finetuned on the task is already sufficient to significantly improve the quality of generated programs across multiple programming languages.

Based on this finding, we propose SLMFix, a framework that leverages an LLM pretrained on general code generation tasks to generate an initial version of the program based on the input query and employs an SLM specialized in error-fixing for the target language to fix statically detected errors in the generated program. We finetune the SLM using reinforcement learning (RL) techniques because performing supervised fine-tuning (SFT) on next-token prediction task fails to address the importance of generating syntactically correct programs. To ensure that the output of the SLM fixes the errors while aligning to the original prompt, we set the reward to be a weighted combination of a static

validator and a semantic scorer, which verify syntactic and other statically checkable correctness properties and assess the functional correctness by comparing to a ground truth program, respectively.

Existing approaches (Zhang et al., 2023; Le et al., 2022; Orlanski et al., 2023) usually evaluate functional correctness of the generated program through sets of test cases. However, constructing such test suites requires significant effort and is necessarily far from complete since the range of valid programs and input data values for those programs is nearly unbounded; additionally, executing the generated programs could introduce additional risks that must be mitigated. Previous studies (Liang et al., 2025; Song et al., 2024) have shown that comparing abstract syntax trees (ASTs) could be an effective workaround for checking the functional correctness without the presence of a test suite, by parsing the generated program and the corresponding ground-truth program into ASTs and comparing the similarities of the ASTs. We believe that the task-specific and well-structured nature of many DSLs means that they have a limited number of programs that achieve the same functionality and that this means they are especially well suited to such an AST-based approach. By comparing AST similarity scores and the Execution Match, we found that AST similarity metric is able to correctly predict the execution result of over 75% of the data points, showing that AST similarity is able to provide an accurate estimate of the functional correctness of the generated program.

Therefore, we propose to use the AST similarity between the output program and ground truth as the semantic score for the reward function, while the static validation can be performed by a variety of static tools, from parsers, to type-checkers, or even symbolic interpreters. The training dataset for the finetuning is generated by first manually crafting a very small dataset with a good coverage of the target language containing a matching set of natural language query and ground truth program. We then employ LLMs to generate corresponding (potentially incorrect) programs as examples for error-fixing.

We evaluate our proposed approach on three example DSLs: Ansible, Bash, and SQL. While Bash and SQL are high-resource languages with large scale training data, Ansible, a popular IT automation tool developed by RedHat, remains a low-resource language due to the specialized nature of its YAML playbooks and limited dataset availability. To verify the effectiveness of our method on Ansible, we constructed an Ansible dataset consisting of Ansible playbooks scraped from open-source repositories on Ansible Galaxy¹. With comprehensive experiments on multiple LLMs and SLMs, we showed that SLMFix is able to successfully eliminate most static errors in the generated program and brings

¹<https://galaxy.ansible.com/>

substantial improvements to the base models across multiple DSLs. In particular, SLMFix outperforms supervised finetuning approach and other strong baselines on Ansible even for very strong code LLMs, which produce significant errors due to the lack of training data, demonstrating the effectiveness of our approach for LRPLs.

In summary, our contributions are as follows:

- We proposed SLMFix, a novel code generation pipeline for DSLs that leverages a pretrained LLM for initial code generation and a SLM fine-tuned using RL for error-fixing.
- We verified the effectiveness of using semantic similarity based on ASTs to evaluate the functional correctness of the LLM-generated programs in DSLs that can work as an alternative for a well-designed test suite.
- We constructed a new dataset for Ansible code generation, scraped from open-source repositories on Ansible Galaxy, that can be utilized for Ansible program generation training and evaluation.
- Our experimental results show that SLMFix is able to significantly improve the quality of LLM-generated programs across multiple DSLs, including both high-resource and low-resource languages, which demonstrates the effectiveness of leveraging SLMs for program repair.

2 RELATED WORK

2.1 Reinforcement Learning for Code Generation

Following the success of Reinforcement Learning (RL) for LLM training, both academic researchers and industry leading companies have started to explore RL approaches in training code generation LLMs. Unlike supervised finetuning (SFT) with the objective of next token prediction, RL methods also consider the syntactic and functional correctness by leveraging compilers and test suites to evaluate the generated code, and are thus more suitable for training LLMs for code generation. CodeRL (Le et al., 2022) proposed to train a separate LLM to predict the functional correctness of the generated program with the unit test results and the solution program. Following works further improved this approach by introducing more fine-grained reward functions, such as incorporating syntactic and semantic matching scores into the reward function (Shojaee et al., 2023), masking unexecuted code segments for more precise model optimization (Dou et al., 2024), designing different types of feedback based on unit test results (Liu et al., 2023), combining symbolic feedback from a formal interpreter with unit tests (Jha et al., 2024), and providing

step-level feedback for multi-step reasoning process during generation (Ye et al., 2025).

However, these RL approaches usually require a test suite to verify the functional correctness of the code. To address this issue, CompCoder (Wang et al., 2022) introduced a multi-stage training pipeline that involves supervised finetuning, reinforcement learning from compiler feedback, and discrimination training that further improves the LLM’s capability in fixing compilation errors, while RLCF (Jain et al., 2023) explored employing an additional LLM to distinguish between the ground-truth and the generated program, using the success rate as part of the reward. These RL-based methods are primarily designed for HRPLs due to the demand for large scale training dataset, and are thus not applicable for LRPLs.

2.2 Automated Error Fixing for LLM-generated Code

Automated error fixing leverages LLMs to diagnose and fix the errors in the code and is commonly employed in the code generation pipelines to improve the quality of LLM-generated code. One common approach to this line of research is training the LLM for error fixing. Self-Edit (Zhang et al., 2023) created a dataset from programs generated by LLMs and train an editor that refines the code based on testing feedback with the dataset. FastFixer (Liu et al., 2024) adopted a similar supervised finetuning approach but also employed a masking strategy to help the model locate the code snippets to fix. Some works have also explored techniques for error fixing that do not require training the models, since training is quite expensive, such as Ngassom et al. (2024) who generate targeted verification questions based on common bug patterns to provide guidance for the LLM to fix the incorrect program without requiring training on the task and MGDebugger (Shi et al., 2024) who propose to isolate bugs at different levels of granularity to identify and fix them more efficiently. In summary, most automated error fixing works focused on supervised finetuning and training-free approaches, but few works have explored leveraging reinforcement learning methods for training the error-fixing model.

Small language models (SLMs) have received plenty of attention from the AI community recently because their limited model size allows for more efficient training and inference. Several prior works have explored the possibility of using SLMs for code generation and related tasks. Sheng & Xu (2025) showed that SLM with less than 1B parameters can still perform well on SQL code generation with well-designed training method. Meanwhile, Kusama et al. (2025) and Koutcheme et al. (2024) benchmarked several SLMs and verified that that SLMs with 3B parameters can achieve comparable performance or even outperform state-of-the-art LLMs with over 10B parameters on program

repair tasks. However, few studies have explored methods leveraging SLMs to fix errors in programs generated by LLMs to improve the quality of the generated programs.

3 METHOD

Our proposed training and inference pipelines are summarized in Figure 1. For inference, our framework uses a frozen LLM pretrained for general code generation and an SLM finetuned for program repair of the target language. The LLM first generates an initial version of the program given the natural language query, which is then passed into a static validator to detect errors such as syntax and type errors. If the validator reports errors in the program, the program, along with the original prompt and error messages from the validator, are provided to the SLM which generates a revised program as the final output. If the LLM initial program does not have errors it becomes the final output.

In Subsection 3.1, we discuss our proposed training method for the SLM. In Subsection 3.2, we describe the design of the validator we employed in both training and inference. In Subsection 3.3, we discuss our approach to developing semantic similarity metrics which is used as an additional metric in training. Finally, Subsection 3.4 outlines our process for generating large amounts of training data.

3.1 Training

Supervised finetuning (SFT) is the most popular approach to adapt LLMs to a new task. However, the training target of SFT is next token prediction, which is not ideal for code-related tasks because it fails to address the importance of generating syntactically and functionally correct programs. Therefore, we decided to use reinforcement learning (RL) with reward signals computed from the feedback of a static validator, which checks properties like syntax and types, and the score given by a semantic similarity metric, which compares the generated program with the ground truth. The static validator generates a boolean signal indicating if the given program contains errors, while the semantic metric will output a score between 0 and 1 representing how well the program aligns with the input query. The reward of the program is then a weighted average between these two signals, and the model parameters will be updated accordingly with proximal policy optimization (PPO) (Schulman et al., 2017).

In order to balance between static and semantic correctness, we weight the scores from the validator and semantic metric in the reward function adaptively based on the percent of generated programs, in the current batch, that pass the validator. With this, the training focuses on improving static validity when most fixed programs still contain errors, and starts to focus on semantic correctness when fixed programs

have a high pass rate on the validator. More formally, given a batch of programs p_1, \dots, p_n , the reward given to program p_i would be

$$r_i = (1 - pr) \cdot \mathbb{1}[f_s(p_i) = \text{true}] + pr \cdot f_f(p_i)$$

where $f_s(p) \in \{\text{true}, \text{false}\}$ is the response of the static validator, $f_f(p) \in [0, 1]$ evaluates the semantic similarity between the program and the ground-truth, and $pr = \frac{1}{n} \sum [f_s(p_i) = \text{true}]$ is the pass rate of the current batch.

3.2 Validator Design

3.2.1 Ansible

For validating Ansible playbooks, we use the Ansible symbolic interpreter developed by Councilman et al. (2025). This system includes a parser for (a subset of) Ansible and thereby validates the syntax of generated programs, and also performs type-checking of the Ansible program, including ensuring variables are well-defined and that module argument are well typed, and it ensures that arguments to module are appropriate (i.e., ensures that required arguments or mutually-exclusive arguments are provided appropriately). Additionally, the interpreter performs symbolic interpretation of the Ansible program which simulates the behavior of the program on all different possible inputs. In this work, we simply use the interpreter for static checking, as described already, but the interpreter could also be used to run tests or to verify the behavior of the program matches some other reference, such as a formal query (Councilman et al., 2025).

3.2.2 Bash

To check the syntactic correctness for Bash, we leveraged ShellCheck (Holen, 2025) tool to check for syntactic errors in the Bash script. Since the tool will also check for check-style errors and linting errors, we only consider the errors that would cause the Bash script to fail the compilation, while omitting other types of errors reported by ShellCheck.

3.2.3 SQL

For SQL, we use SQLGlot (Mao, 2025) to detect syntactic errors in the query. We also employ sql-metadata (Bencz, 2025) library to extract all the table and column names referenced in the SQL query, and check if all referenced names are present in the provided database schema to further verify the correctness of the query.

3.3 Semantic Similarity Metric Design

Domain-specific languages (DSLs) are typically task-specific, well-structured, and highly abstract languages. Therefore, the commands in DSLs are high-level and usually tailored to the specific domain that the language applies to, greatly limiting the number of different ways to com-

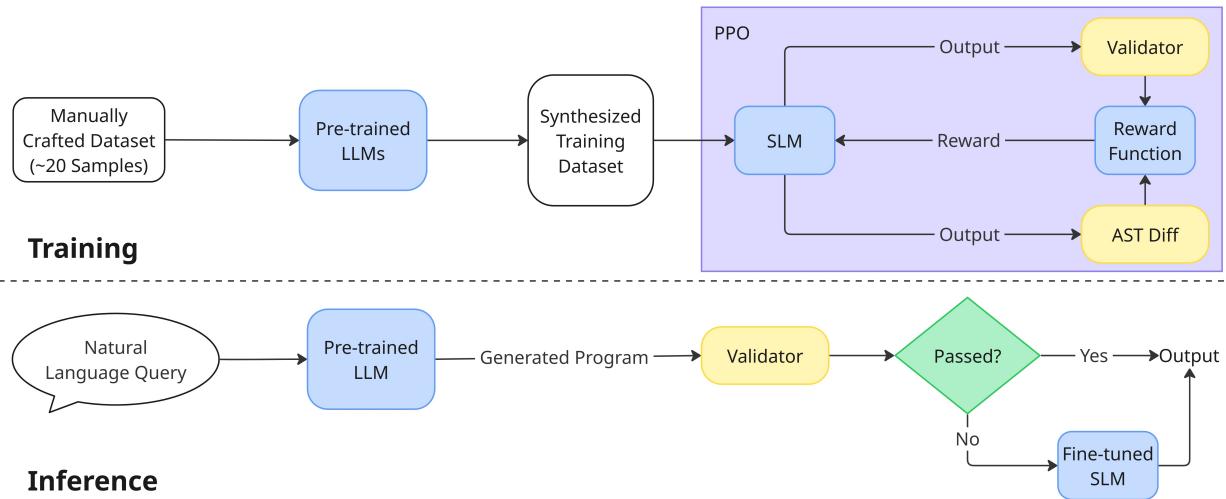


Figure 1. The overview of the proposed training and inference pipelines.

plete a specific task. Based on this characteristic of DSLs, we proposed to measure how well the generated program aligns with the input query by comparing the corresponding abstract syntax trees (ASTs) of the generated program and a reference ground-truth program. ASTs are the result of parsing the target language program and often abstract some syntactic details, meaning two syntactically different programs may have the same or very similar ASTs. The following subsections describe the language-specific AST-based semantic similarity metric we design for the programming languages in our evaluation.

3.3.1 Ansible

In this work, we focus on generating Ansible playbooks, which define automation pipelines in YAML format that Ansible uses to configure and manage nodes. An Ansible playbook contains a sequence of **plays** that define the operations to perform. Each play consists of a sequence of **tasks** that define the atomic operations in the play, as well as other configurations that specify how the tasks should be executed.

Following the practice of [Pujar et al. \(2023\)](#), we decompose the generated playbook into a sequence of plays, and further divide each play into a list of tasks, repeating this procedure for the ground-truth. We parse each task into a dictionary, where the keys are arguments provided to the task (the optional name field is omitted). For each task in the ground-truth, we find the corresponding task in the outputting playbook and check if each key in the task dictionary appears in the corresponding task dictionary of output and if the same value is assigned to the key. The final score

assigned to a particular task is

$$score = \frac{\sum_{(k,v) \in gt} \mathbb{1}[(k,v) \in pred]}{|gt|}$$

where gt is the task dictionary of ground-truth and $pred$ is the task dictionary of prediction. We then obtain the play-level score by averaging over all tasks in the play, and eventually compute the score for the whole playbook as the average score of all plays in the playbook.

3.3.2 Bash

Since there are no existing datasets for multi-line Bash script generation, we only focus on parsing single-line Bash scripts that might include parentheses and connectors. We use Bashlex ([Kamara, 2025](#)) to parse the Bash script into *atomic commands*, where each atomic command executes a single bash command without any connectors. Similar to Ansible, we parse each atomic command into a dictionary, where each key is an option and the value is its argument, and check if each key in the ground-truth appears in the output and is associated with the same value. We obtain the overall score for the Bash script by taking the average over all its commands.

3.3.3 SQL

For SQL, we use SQLGlot ([Mao, 2025](#)) to normalize the query and parse it into an AST. To improve the accuracy of the AST-based comparison, we also remove all column aliases and replace table aliases with their original names. The semantic similarity score is computed as the reciprocal of the number edits required to convert the AST of the generated program to the AST of the ground-truth.

3.4 Training Data Synthesis

To construct training data for the reinforcement learning from symbolic feedback (RLSF) phase, we adopted a multi-model generation strategy that leverages the diversity of outputs from different large language models. From each domain-specific dataset (Ansible, SQL, and Bash), we manually compiled a subset of 20 natural language queries along with their corresponding ground truth code implementations, language’s core constructs, syntactic patterns, and common programming idioms. With this compact set of examples, we sampled 50 sample programs for each example from five strong code LLMs: StarCoder-2 7B (Lozhkov et al., 2024), DeepSeek-Coder 6.7B (Guo et al., 2024), Qwen2.5-Coder 7B (Hui et al., 2024), Granite 8B Code (Mishra et al., 2024), and LLaMA-3.1 8B (LlamaTeam & AI@Meta, 2024).

The motivation for this approach stems from research demonstrating that ensemble methods and multi-model code generation strategies can significantly improve code quality and robustness. By soliciting code generations from multiple LLMs with varying architectures, training procedures, and capabilities, we obtain diverse candidate implementations that capture different problem-solving strategies and programming patterns. Each LLM in the ensemble contributes its unique strengths while potentially avoiding the systematic errors or biases present in individual models.

This multi-model generation process creates an expansive training dataset consisting of natural language queries paired with both LLM-generated code (of varying quality) and corresponding ground truth implementations. The resulting dataset provides the varied examples necessary for training a reward model that can distinguish between high-quality and low-quality code generations, which is essential for the RLSF training procedure. The preference pairs created by comparing LLM-generated code against ground truth enable the reward model to learn human preferences regarding code correctness, style, and efficiency.

The remainder of the collected data from each domain (beyond the 20 samples used for RLSF training) was reserved for validation and evaluation purposes. This partitioning strategy ensures that model performance is assessed on held-out examples that were not involved in either the supervised finetuning or reinforcement learning phases, providing an unbiased estimate of generalization capability to novel natural language-to-code translation tasks.

4 ANSIBLE DATASET CONSTRUCTION

This section describes the comprehensive methodology employed to construct the dataset for Ansible code generation. The Ansible dataset was constructed through an end-to-end collection and parsing pipeline designed to gather real-world infrastructure automation code. The data collection process

began with the development of a custom web scraper built using Selenium, a widely-used browser automation framework that enables interaction with dynamic web content. This scraper systematically crawled the Ansible Galaxy website, the primary community hub for sharing Ansible collections and roles, to identify and download all publicly available repositories listed on the platform.

Following the repository collection phase, the downloaded content underwent structured parsing using the Ansible Content Parser (Ansible, 2025a) tool. This specialized parser, which operates by running Ansible Lint (Ansible, 2025b) internally, was designed specifically to analyze Ansible files and extract valid, parseable playbooks from the collected repositories. The Ansible Content Parser examines the structure of Ansible content including playbooks, roles, and tasks, identifying syntactically correct playbooks while filtering out malformed or unparsable content.

A critical component of the parsing process involved variable and role infilling, which addresses the challenge that Ansible playbooks frequently contain placeholder variables defined using Jinja2 templating syntax and imported roles defined elsewhere in the repository. For variable infilling, we used the parser to identify variable references within playbooks and populated them with appropriate values based on variable definitions found in defaults, vars files, inventory configurations, and other sources following Ansible’s variable precedence rules. For role infilling, we collected all roles that are defined in the scope of the playbook and replace the roles by their corresponding definitions in the playbook. This variable and role resolution process transforms abstract, template-based playbooks into concrete, executable examples suitable for training machine learning models.

The web scraping and parsing pipeline successfully yielded a substantial corpus of syntactically valid Ansible playbooks with resolved variables. However, these playbooks lacked accompanying natural language descriptions that would be necessary for training natural language-to-code generation tasks. To address this limitation, we employed GPT-4o (OpenAI et al., 2024), a state-of-the-art large language model optimized for natural language processing tasks with reduced computational requirements. GPT-4o was utilized to automatically generate natural language prompts for each valid playbook in the collected dataset. This synthetic prompt generation process leverages the model’s extensive pretraining on diverse text corpora and its demonstrated capability to understand and describe code functionality, transforming the code-only corpus into paired natural language-code examples suitable for supervised learning.

	Ansible				Bash				SQL			
	BLEU	CodeBERTScore	Pass Rate	AST Diff	BLEU	CodeBERTScore	Pass Rate	AST Diff	BLEU	CodeBERTScore	Pass Rate	AST Diff
Qwen-2.5-Coder 7B	Base	0.4063	0.8194	59.40%	0.3100	0.5284	0.8698	98.80%	0.4975	0.6477	0.9072	97.78%
	SFT	0.4643	<u>0.8376</u>	71.82%	0.3563	<u>0.6095</u>	<u>0.8904</u>	99.66%	<u>0.5503</u>	0.8138	<u>0.9536</u>	77.85%
	ICL	0.4275	0.8386	65.60%	0.3390	0.5557	0.8556	99.32%	0.5143	0.7322	0.9396	97.68%
	Self-correction	0.4072	0.8201	60.74%	0.3147	0.5284	0.8699	99.15%	0.4988	0.6478	0.9073	98.36%
	Self>Edit	0.3651	0.7623	61.41%	0.2825	0.5269	0.7851	99.66%	0.4970	0.6389	0.8701	97.78%
	SLMFix(Ours)	0.4435	0.8304	97.82%	0.3974	0.5310	0.8459	<u>99.83%</u>	0.5012	0.6458	0.8866	<u>99.71%</u>
DeepSeek-Coder 6.7B	Base	0.4030	0.8208	60.23%	0.3317	0.4800	0.8516	100%	0.4349	0.5562	0.8783	99.71%
	SFT	0.3714	0.8073	44.97%	0.3081	<u>0.5898</u>	<u>0.8816</u>	97.61%	<u>0.5247</u>	0.7009	0.9180	83.08%
	ICL	0.4353	<u>0.8327</u>	60.74%	0.3389	0.5034	0.8563	99.66%	0.4445	0.5773	0.8859	97.87%
	Self-correction	0.4031	0.8186	63.09%	0.3321	0.4800	0.8516	100%	0.4349	0.5563	0.8781	99.90%
	Self>Edit	0.3748	0.7679	63.76%	0.2883	0.4800	0.8516	100%	0.4349	0.5563	0.8726	99.71%
	SLMFix(Ours)	0.4530	0.8302	97.99%	<u>0.3994</u>	0.4800	0.8516	100%	0.4349	0.5562	0.8846	100%
StarCoder2 7B	Base	0.3020	0.7958	55.37%	0.2069	0.5149	0.8644	99.32%	0.4664	0.5515	0.8712	99.03%
	SFT	0.4547	<u>0.8357</u>	89.60%	<u>0.3382</u>	0.6209	<u>0.8961</u>	97.61%	<u>0.5398</u>	<u>0.6924</u>	<u>0.9189</u>	71.28%
	ICL	0.3897	0.8238	67.95%	0.3130	0.5034	0.8563	99.66%	0.4445	0.6162	0.8989	97.00%
	Self-correction	0.3076	0.7979	57.72%	0.2053	0.5149	0.8638	99.49%	0.4664	0.5514	0.8711	99.23%
	Self>Edit	0.3034	0.7579	60.57%	0.2138	0.5158	0.7888	99.83%	0.4664	0.5504	0.8626	99.03%
	SLMFix(Ours)	0.3837	0.8265	<u>95.64%</u>	0.3201	0.5163	0.8424	100%	0.4687	0.5541	0.8852	<u>99.81%</u>

Table 1. Evaluation results of employing LLMs as base models on Ansible, Bash, and SQL code generation. The best performance in each column is marked in **bold**, while the best performance in each group is marked in underline. SLMFix is able to achieve the highest pass rate on the validator across all languages and LLMs, achieving competitive performance and even outperforms the strongest baselines on semantic metrics.

5 EXPERIMENTS

5.1 Experiment Setup

Target Programming Languages. We choose Ansible, Bash, and SQL as the examples of domain-specific languages (DSLs) in our experiment. Ansible, a popular IT automation tool developed by RedHat, is a low-resource programming language (LRPLs) despite being critical in many fields. On the other hand, Bash and SQL are examples of high-resource DSLs with abundant training data. We include both low-resource and high resource languages to demonstrate the effectiveness of our approach across diverse DSLs.

Datasets. For Ansible code generation, we use the dataset we constructed, as described in Section 4. For SQL command generation, we leveraged the Spider dataset (Yu et al., 2018), a large-scale human-labeled corpus designed specifically for complex and cross-domain semantic parsing and text-to-SQL tasks. The dataset consists of 10,181 natural language questions paired with 5,693 unique complex SQL queries spanning 200 databases with multiple tables across 138 different domains. To prepare the Spider data for model training, the natural language queries were augmented with their corresponding database schemas, which contains structured metadata describing the tables, columns, data types, and foreign key relationships. By appending this schema information to each natural language query, the dataset provides models with the necessary context to understand the database structure and generate syntactically correct SQL queries that reference valid table and column names. The Bash dataset used was the NL2Bash dataset

(Lin et al., 2018), a corpus containing approximately 10,000 English descriptions paired with their corresponding Bash one-liners. The corpus covers over 100 commonly used Bash utilities and includes diverse commands involving file manipulation, searching, piping, and system operations.

Baselines. We selected Qwen-2.5-Coder 7B (Hui et al., 2024), DeepSeek-Coder 6.7B (Guo et al., 2024), and StarCoder2 7B (Lozhkov et al., 2024) models as the base LLMs and compared our proposed approach against the following types of baselines in our experiments:

- Pre-trained LLMs: directly generate the program given the input query;
- Supervised fine-tuning (SFT): finetune the base LLM on target language and employ finetuned LLMs for direct code generation;
- In-context learning (ICL): provide several examples of the target language to the base LLM for direct code generation;
- Self-correction: provide the base LLM with the error message from the validator and ask it to revise the original program;
- Automatic program repair methods: finetune the SLM for error-fixing with Self>Edit (Zhang et al., 2023), the example of automatic program repair methods we selected.

Since these pretrained LLMs are already able to achieve very high pass rates on the validators for Bash and SQL, we also

	Ansible				Bash				SQL			
	BLEU	CodeBERTScore	Pass Rate	AST Diff	BLEU	CodeBERTScore	Pass Rate	AST Diff	BLEU	CodeBERTScore	Pass Rate	AST Diff
Qwen-2.5 0.5B												
Base	0.2689	0.7695	19.63%	0.2006	0.4352	<u>0.8334</u>	97.09%	0.3961	0.4242	0.8384	64.41%	0.2091
SFT	0.2215	0.7465	10.57%	0.1014	0.2258	0.7508	90.26%	0.1886	<u>0.5308</u>	0.8722	51.55%	0.2556
ICL	0.3038	0.7879	31.21%	0.1868	0.4268	0.8308	96.75%	0.3691	0.4617	0.8520	56.57%	0.1953
Self-correction	0.2715	0.7732	20.13%	0.2017	0.4352	0.8333	97.44%	0.3961	0.4244	0.8384	64.51%	0.2091
Self>Edit	0.2161	0.7392	30.20%	0.1836	0.4353	0.7801	<u>99.32%</u>	0.3938	0.3866	0.8107	64.60%	0.1886
SLMFix(Ours)	<u>0.3913</u>	<u>0.8274</u>	<u>94.30%</u>	<u>0.3455</u>	<u>0.4372</u>	0.8324	<u>99.32%</u>	<u>0.3984</u>	0.4679	<u>0.8777</u>	<u>93.71%</u>	<u>0.2789</u>
LLaMA-3.2 1B												
Base	0.2931	0.7770	18.96%	0.2328	0.4137	0.8310	97.09%	0.3933	0.5384	0.8787	85.20%	0.3547
SFT	0.3899	0.8036	52.35%	0.2428	<u>0.4217</u>	0.8188	93.16%	0.4000	<u>0.6395</u>	0.9067	59.09%	0.3471
ICL	0.3451	0.7967	33.22%	0.2508	0.4195	0.8286	98.12%	0.3939	0.5739	0.8872	73.89%	0.3501
Self-correction	0.2993	0.7796	21.31%	0.2378	0.4142	0.8172	97.78%	0.3950	0.5362	0.8727	85.40%	0.3540
Self>Edit	0.2419	0.7402	24.33%	0.1586	0.4139	0.7805	<u>99.49%</u>	0.3933	0.5299	0.8524	85.20%	0.3470
SLMFix(Ours)	<u>0.4104</u>	<u>0.8290</u>	<u>92.79%</u>	<u>0.3558</u>	0.4186	<u>0.8332</u>	<u>99.49%</u>	<u>0.4001</u>	0.5451	0.8832	<u>97.20%</u>	<u>0.3763</u>
DeepSeek-Coder 1.3B												
Base	0.3477	0.7987	39.60%	0.2201	0.4454	0.8381	98.12%	0.3971	0.5160	0.8643	96.23%	0.4029
SFT	0.3876	0.8107	56.38%	0.2771	<u>0.5308</u>	<u>0.8621</u>	<u>98.12%</u>	<u>0.4635</u>	<u>0.6583</u>	<u>0.9114</u>	73.50%	0.4306
ICL	0.3827	0.8091	41.78%	0.2252	0.4723	0.8440	98.97%	0.4212	0.5584	0.8820	95.36%	0.4169
Self-correction	0.3458	0.7976	42.62%	0.2239	0.4453	0.8376	99.83%	0.3971	0.5163	0.8633	96.42%	0.4031
Self>Edit	0.3035	0.7550	47.15%	0.2211	0.4462	0.7838	<u>100%</u>	0.3971	0.5131	0.8547	96.23%	0.4003
SLMFix(Ours)	<u>0.4144</u>	<u>0.8304</u>	<u>95.97%</u>	<u>0.3384</u>	0.4487	0.8346	<u>100%</u>	0.4019	0.5175	0.8837	<u>99.32%</u>	<u>0.4063</u>
Granite-3.3 2B												
Base	0.3828	0.8103	43.79%	0.2864	0.4432	0.8385	97.98%	0.4056	0.4943	0.8581	89.17%	0.3423
SFT	0.4257	0.8217	63.59%	0.2952	<u>0.5621</u>	<u>0.8713</u>	<u>98.46%</u>	<u>0.4830</u>	<u>0.6899</u>	<u>0.9184</u>	72.05%	<u>0.4160</u>
ICL	0.3954	0.8173	48.15%	0.3020	0.4754	0.8486	97.61%	0.4362	0.5351	0.8708	89.56%	0.3744
Self-correction	0.3794	0.8019	48.83%	0.2843	0.4428	0.8394	99.15%	0.4124	0.4948	0.8566	91.97%	0.3457
Self>Edit	0.3303	0.7563	47.99%	0.2383	0.4400	0.7806	99.49%	0.4060	0.4894	0.8455	89.17%	0.3361
SLMFix(Ours)	<u>0.4344</u>	<u>0.8310</u>	<u>96.64%</u>	<u>0.3782</u>	0.4412	0.8358	<u>99.66%</u>	0.4174	0.5022	0.8840	<u>98.65%</u>	0.3598

Table 2. Evaluation results of employing SLMs as base models on Ansible, Bash, and SQL code generation. The best performance in each column is marked in **bold**, while the best performance in each group is marked in underline. SLMFix is able to achieve the highest pass rate on the validator across all languages and SLMs, outperforming the strongest baselines on semantic metrics.

selected smaller LLMs, including Qwen-2.5 0.5B (Yang et al., 2024), LLaMA-3.2 1B (LlamaTeam & AI@Meta, 2024), DeepSeek-Coder 1.3B (Guo et al., 2024), and Granite-3.3 2B (Mishra et al., 2024), to evaluate the performance of our approach when the base LLMs are less strong. For our error-fixing model, we selected Qwen-2.5 Coder 0.5B (Hui et al., 2024) which we fine-tune for the task. For in-context learning, we randomly selected five samples from the training dataset as the in-context examples.

Metrics. For all languages, we report BLEU score (Papineni et al., 2002) and CodeBERTScore (Zhou et al., 2023) to evaluate the semantic similarity between the generated program and the ground-truth. For each language, we also report the pass rate on the corresponding validator and the average score on the semantic similarity metric described in Subsection 3.3.

5.2 Training Details

For reinforcement learning approaches, we use verl (Sheng et al., 2024) and vLLM (Kwon et al., 2023) to perform training with 70 total epochs and batch size of 64 on the Qwen-2.5-Coder 0.5B model. We set the maximum input and output length to be 2048 tokens, and the learning rates for actor model and critic model as 1e-6 and 1e-5, respectively. For supervised finetuning, we employed LoRA (Hu et al., 2022) for parameter-efficient finetuning. We set the maximum number of steps to be 150 and batch size as 4,

with learning rate of 1e-4 and weight decay of 0.05. At inference time, we use vLLM for efficient generation, setting temperature as 0 to avoid randomness.

5.3 Main Results

LLMs as the base model. Table 1 presents the evaluation results of SLMFix and baselines on Ansible, Bash, and SQL code generation using LLMs as base models. SLMFix is able to achieve the highest pass rate on the validator across all models and languages, showing its strong capability in eliminating static errors in LLM-generated program. In addition, SLMFix also demonstrates that it is capable of improving the functional correctness of generated programs at the same time, especially for Ansible, improving the correctness of LLM-generated programs substantially and outperforming SFT by 4 points on Qwen-2.5-Coder 7B and 9 points on DeepSeek-Coder 6.7B.

SLMs as the base model. Table 2 presents the evaluation results of SLMFix and baselines on Ansible, Bash, and SQL code generation using SLMs as base models. Similarly, SLMFix receives the highest pass rate on the validator across all models and languages, achieving comparable performance or even outperforming SFT baselines. In particular, SLMFix outperforms all baseline methods in terms of BLEU Score, CodeBERTScore, and AST Diff metrics for Ansible code generation, demonstrating its effectiveness in improving the performance of SLMs in low-resource

languages.

5.4 Comparison between AST Diff and Execution Match

To verify the effectiveness of our proposed AST Diff metric, we compared the score given by AST Diff with Execution Match on the training set of Spider dataset (Yu et al., 2018) with test suites provided by Zhong et al. (2020). We selected Qwen-2.5-Coder 7B, DeepSeek-Coder 6.7B, and StarCoder2 7B, and computed the accuracy of AST Diff using Execution Match as the ground-truth. In specific, we check if two SQL queries that are equivalent in AST Diff leads to the same execution results and vice versa. The results are summarized in Figure 3. The results demonstrate that the AST Diff metric is able to predict an average of more than 75% of the programs correctly with very low false positive (FP) rate.

	TP	FP	FN	TN	Acc
Qwen-2.5-Coder 7B	417	101	179	337	72.92%
DeepSeek-Coder 6.7B	391	7	217	419	78.34%
StarCoder2 7B	362	14	234	424	76.02%

Table 3. Accuracy of AST Diff metric using Execution Match as ground-truth. We computed the number of true positive (TP), false positive (FP), false negative (FN), and true negative (TN), and also report the accuracy (Acc) of the metric. The AST Diff metric is able to achieve an average of above 75% across the models in comparison to Execution Match.

5.5 Ablation Studies

	Ansible			
	BLEU	CodeBERTScore	Pass Rate	AST Diff
RL for code generation	0.3856	0.8101	48.83%	0.2792
Qwen-2.5 0.5B				
SLMFix	<u>0.3913</u>	0.8274	94.30%	0.3455
Syntactic score only	0.3265	<u>0.8298</u>	<u>98.99%</u>	0.2062
Semantic score only	0.2553	0.7588	22.48%	0.4047
LLaMA-3.2 1B				
SLMFix	0.4104	0.8290	92.79%	0.3558
Syntactic score only	0.3612	0.8365	99.50%	0.2353
Semantic score only	0.2462	0.7430	19.63%	<u>0.3601</u>

Table 4. The results of ablation studies. The best performance in each column is marked in **bold**, while the best performance in each group is marked in underline. Our proposed training method can achieve an excellent balance between the syntactic and functional correctness compared to the ablated settings.

To better understand how different components in the training approach affect the overall performance of our proposed framework, we conducted a series of ablation studies on the training settings. To highlight the performance difference across various settings, we select Ansible as the target language and employ Qwen-2.5 0.5B and LLaMA-3.2 1B

as the base model to make the program repair task more challenging for SLMFix. Table 4 summarizes the results of the ablation study. We found that using only the static validator score or the semantic similarity metric would achieve excellent performance on the particular metric the model was trained on, but also cause the trained model to overfit to the training target while ignoring other metrics, resulting in serious performance degradation. Meanwhile, combining both the static validator score and the semantic similarity metric achieves a better balance between static validation and functional correctness, helping the model to generate higher quality programs. In addition, we found that the reinforcement learning method we applied to train the SLM for error-fixing task does not work well on direct code generation task, possibly due to the limited model capability of SLMs. This result demonstrate the benefits of leveraging SLMs for program repair over direct generation.

6 CONCLUSION

In this paper, we present SLMFix, a new pipeline for code generation that involves a pretrained LLM and an SLM tailored for error-fixing. We propose to leverage deep reinforcement learning techniques to train the small language model to fix errors identified in the program generated by the LLM, using a fine-grained reward function that encourages both static and functional correctness of the modified program. We believe this approach is especially well suited to DSLs as it leverages the power of reinforcement learning while using a small model that is less resource intensive and requires less data to train. We also verify the effectiveness of using similarity of abstract syntax trees as an estimate of functional correctness for DSLs to replace the need for a well-designed test suite. With extensive experiment on Ansible, Bash, and SQL, we demonstrated that our proposed method is able to significantly improve the quality of programs generated by LLMs and outperforms similar program repair approaches, achieving comparable performance with LLM finetuned on the task.

ACKNOWLEDGEMENTS

This work is supported by the IBM-ILLINOIS Discovery Accelerator Institute (IIDAI). Any opinions, findings, or conclusions expressed in this material are those of the authors and do not necessarily reflect the views of IBM.

This research used the Delta advanced computing and data resource which is supported by the National Science Foundation (award OAC 2005572) and the State of Illinois. Delta is a joint effort of the University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications. This use was through allocation CIS250276 from the Advanced Cyberinfrastructure Coordination Ecosystem: Ser-

vices & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

REFERENCES

- Ansible. ansible/ansible-content-parser, May 2025a. URL <https://github.com/ansible/ansible-content-parser>. original-date: 2023-07-20T19:02:08Z.
- Ansible. ansible/ansible-lint, October 2025b. URL <https://github.com/ansible/ansible-lint>. original-date: 2013-08-14T11:08:00Z.
- Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.
- Brencz, M. macbre/sql-metadata, October 2025. URL <https://github.com/macbre/sql-metadata>. original-date: 2017-06-06T15:59:09Z.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code, 2021. URL <https://arxiv.org/abs/2107.03374>.
- Councilman, A., Fu, D., Gupta, A., Wang, C., Grove, D., Wang, Y.-X., and Adve, V. Towards formal verification of llm-generated code from natural language prompts, 2025. URL <https://arxiv.org/abs/2507.13290>.
- Dou, S., Liu, Y., Jia, H., Zhou, E., Xiong, L., Shan, J., Huang, C., Wang, X., Fan, X., Xi, Z., Zhou, Y., Ji, T., Zheng, R., Zhang, Q., Gui, T., and Huang, X. Step-Coder: Improving Code Generation with Reinforcement Learning from Compiler Feedback. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pp. 4571–4585. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.251. URL <https://doi.org/10.18653/v1/2024.acl-long.251>.
- Gemini Team, G. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL <https://arxiv.org/abs/2507.06261>.
- Gong, Z., Guo, Y., Zhou, P., Gao, C., Wang, Y., and Xu, Z. MultiCoder: Multi-Programming-Lingual Pre-Training for Low-Resource Code Completion. *CoRR*, abs/2212.09666, 2022. doi: 10.48550/ARXIV.2212.09666. URL <https://doi.org/10.48550/arXiv.2212.09666>. arXiv: 2212.09666.
- Guo, D., Zhu, Q., Yang, D., Xie, Z., Dong, K., Zhang, W., Chen, G., Bi, X., Wu, Y., Li, Y. K., Luo, F., Xiong, Y., and Liang, W. DeepSeek-Coder: When the Large Language Model Meets Programming - The Rise of Code Intelligence. *CoRR*, abs/2401.14196, 2024. doi: 10.48550/ARXIV.2401.14196. URL <https://doi.org/10.48550/arXiv.2401.14196>. arXiv: 2401.14196.
- Holen, V. koalaman/shellcheck, October 2025. URL <https://github.com/koalaman/shellcheck>. original-date: 2012-11-17T03:15:11Z.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL <https://openreview.net/forum?id=nZeVKeFYf9>.
- Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., Liu, T., Zhang, J., Yu, B., Dang, K., Yang, A., Men, R., Huang, F., Ren, X., Ren, X., Zhou, J., and Lin, J. Qwen2.5-Coder Technical Report. *CoRR*, abs/2409.12186, 2024. doi: 10.48550/ARXIV.2409.12186. URL <https://doi.org/10.48550/arXiv.2409.12186>. arXiv: 2409.12186.
- Jain, A., Adiole, C., Chaudhuri, S., Reps, T. W., and Jermaine, C. Tuning Models of Code with Compiler-Generated Reinforcement Learning Feedback. *CoRR*, abs/2305.18341, 2023. doi: 10.48550/ARXIV.2305.18341. URL <https://doi.org/10.48550/arXiv.2305.18341>. arXiv: 2305.18341.
- Jha, P., Jana, P., Arora, A., and Ganesh, V. RLSF: Reinforcement Learning via Symbolic Feedback. *CoRR*, abs/2405.16661, 2024. doi: 10.48550/ARXIV.2405.16661. URL <https://doi.org/10.48550/arXiv.2405.16661>. arXiv: 2405.16661.

-
- Kamara, I. idank/bashlex, October 2025. URL <https://github.com/idank/bashlex>. original-date: 2014-09-20T06:45:30Z.
- Koutcheme, C., Dainese, N., Sarsa, S., Leinonen, J., Hellas, A., and Denny, P. Benchmarking Educational Program Repair. *CoRR*, abs/2405.05347, 2024. doi: 10.48550/ARXIV.2405.05347. URL <https://doi.org/10.48550/arXiv.2405.05347>. arXiv: 2405.05347.
- Kusama, K., Shu, H., Kondo, M., and Kamei, Y. How Small is Enough? Empirical Evidence of Quantized Small Language Models for Automated Program Repair. *CoRR*, abs/2508.16499, 2025. doi: 10.48550/ARXIV.2508.16499. URL <https://doi.org/10.48550/arXiv.2508.16499>. arXiv: 2508.16499.
- Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- Le, H., Wang, Y., Gotmare, A. D., Savarese, S., and Hoi, S. C.-H. CodeRL: Mastering Code Generation through Pretrained Models and Deep Reinforcement Learning. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. URL http://papers.nips.cc/paper_files/paper/2022/hash/8636419dealaa9fdbd25fc4248e702da4-Abstract.html.
- Li, Y., Choi, D. H., Chung, J., Kushman, N., Schrittweiser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Lago, A. D., Hubert, T., Choy, P., d'Autume, C. d. M., Babuschkin, I., Chen, X., Huang, P.-S., Welbl, J., Gowal, S., Cherepanov, A., Molloy, J., Mankowitz, D. J., Robson, E. S., Kohli, P., Freitas, N. d., Kavukcuoglu, K., and Vinyals, O. Competition-Level Code Generation with AlphaCode. *CoRR*, abs/2203.07814, 2022. doi: 10.48550/ARXIV.2203.07814. URL <https://doi.org/10.48550/arXiv.2203.07814>. arXiv: 2203.07814.
- Liang, Q., Zhang, Z., Sun, Z., Lin, Z., Luo, Q., Xiao, Y., Chen, Y., Zhang, Y., Zhang, H., Zhang, L., Chenbin, C., and Xiong, Y. Grammar-Based Code Representation: Is It a Worthy Pursuit for LLMs? In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 15640–15653. Association for Computational Linguistics, 2025. URL <https://aclanthology.org/2025.findings-acl.807/>.
- Lin, X. V., Wang, C., Zettlemoyer, L., and Ernst, M. D. NL2Bash: A Corpus and Semantic Parser for Natural Language Interface to the Linux Operating System. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Marian, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T. (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA), 2018. URL <http://www.lrec-conf.org/proceedings/lrec2018/summaries/1021.html>.
- Liu, F., Liu, Z., Zhao, Q., Jiang, J., Zhang, L., Sun, Z., Li, G., Li, Z., and Ma, Y. FastFixer: An Efficient and Effective Approach for Repairing Programming Assignments. In Filkov, V., Ray, B., and Zhou, M. (eds.), *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering, ASE 2024, Sacramento, CA, USA, October 27 - November 1, 2024*, pp. 669–680. ACM, 2024. doi: 10.1145/3691620.3695062. URL <https://doi.org/10.1145/3691620.3695062>.
- Liu, J., Zhu, Y., Xiao, K., Fu, Q., Han, X., Yang, W., and Ye, D. RLTF: Reinforcement Learning from Unit Test Feedback. *Trans. Mach. Learn. Res.*, 2023, 2023. URL <https://openreview.net/forum?id=hjYmsV6nXZ>.
- LlamaTeam and AI@Meta. The llama 3 herd of models. 2024. URL <https://arxiv.org/abs/2407.21783>.
- Lozhkov, A., Li, R., Allal, L. B., Cassano, F., Lamy-Poirier, J., Tazi, N., Tang, A., Pykhtin, D., Liu, J., Wei, Y., Liu, T., Tian, M., Kocetkov, D., Zucker, A., Belkada, Y., Wang, Z., Liu, Q., Abulkhanov, D., Paul, I., Li, Z., Li, W.-D., Risdal, M., Li, J., Zhu, J., Zhuo, T. Y., Zheltonozhskii, E., Dade, N. O. O., Yu, W., Krauß, L., Jain, N., Su, Y., He, X., Dey, M., Abati, E., Chai, Y., Muennighoff, N., Tang, X., Oblokulov, M., Akiki, C., Marone, M., Mou, C., Mishra, M., Gu, A., Hui, B., Dao, T., Zebaze, A., Dehaene, O., Patry, N., Xu, C., McAuley, J. J., Hu, H., Scholak, T., Paquet, S., Robinson, J., Anderson, C. J., Chapados, N., and al. e. StarCoder 2 and The Stack v2: The Next Generation. *CoRR*, abs/2402.19173, 2024. doi: 10.48550/ARXIV.2402.19173. URL <https://doi.org/10.48550/arXiv.2402.19173>. arXiv: 2402.19173.
- Mao, T. tobymao/sqlglot, October 2025. URL <https://github.com/tobymao/sqlglot>. original-date: 2021-03-13T05:01:56Z.

-
- Mishra, M., Stallone, M., Zhang, G., Shen, Y., Prasad, A., Soria, A. M., Merler, M., Selvam, P., Surendran, S., Singh, S., Sethi, M., Dang, X.-H., Li, P., Wu, K.-L., Zawad, S., Coleman, A., White, M., Lewis, M., Pavuluri, R., Koyfman, Y., Lublinsky, B., Bayser, M. d., Abdelaziz, I., Basu, K., Agarwal, M., Zhou, Y., Johnson, C., Goyal, A., Patel, H., Shah, S. Y., Zerfos, P., Ludwig, H., Munawar, A., Crouse, M., Kapanipathi, P., Salaria, S., Calio, B., Wen, S., Seelam, S., Belgodere, B., Fonseca, C. A., Singhee, A., Desai, N., Cox, D. D., Puri, R., and Panda, R. Granite Code Models: A Family of Open Foundation Models for Code Intelligence. *CoRR*, abs/2405.04324, 2024. doi: 10.48550/ARXIV.2405.04324. URL <https://doi.org/10.48550/arXiv.2405.04324>. arXiv: 2405.04324.
- Ngassom, S. K., Dakhel, A. M., Tambon, F., and Khomh, F. Chain of Targeted Verification Questions to Improve the Reliability of Code Generated by LLMs. In Adams, B., Zimmermann, T., Ozkaya, I., Lin, D., and Zhang, J. M. (eds.), *Proceedings of the 1st ACM International Conference on AI-Powered Software, AIware 2024, Porto de Galinhas, Brazil, July 15-16, 2024*. ACM, 2024. doi: 10.1145/3664646.3664772. URL <https://doi.org/10.1145/3664646.3664772>.
- OpenAI. GPT-5 System Card, 2025. URL <https://cdn.openai.com/gpt-5-system-card.pdf>.
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgium, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kirov, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- OpenAI, : , El-Kishky, A., Wei, A., Saraiva, A., Minaiev, B., Selsam, D., Dohan, D., Song, F., Lightman, H., Clavera, I., Pachocki, J., Tworek, J., Kuhn, L., Kaiser, L., Chen, M., Schwarzer, M., Rohaninejad, M., McAleese, N., o3 contributors, Mürk, O., Garg, R., Shu, R., Sidor, S., Kosaraju, V., and Zhou, W. Competitive programming with large reasoning models, 2025. URL <https://arxiv.org/abs/2502.06807>.
- Orlanski, G., Xiao, K., Garcia, X., Hui, J., Howland, J., Malmaud, J., Austin, J., Singh, R., and Catasta, M. Measuring the Impact of Programming Language Distribution. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26619–26645. PMLR, 2023. URL <https://proceedings.mlr.press/v202/orlanski23a.html>.

- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pp. 311–318. ACL, 2002. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040/>.
- Pujar, S., Buratti, L., Guo, X., Dupuis, N., Lewis, B. L., Suneja, S., Sood, A., Nalawade, G., Jones, M., Morari, A., and Puri, R. Invited: Automated Code generation for Information Technology Tasks in YAML through Large Language Models. In *60th ACM/IEEE Design Automation Conference, DAC 2023, San Francisco, CA, USA, July 9-13, 2023*, pp. 1–4. IEEE, 2023. doi: 10.1109/DAC56929.2023.10247987. URL <https://doi.org/10.1109/DAC56929.2023.10247987>.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL <https://arxiv.org/abs/1707.06347>.
- Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.
- Sheng, L. and Xu, S.-S. SLM-SQL: An Exploration of Small Language Models for Text-to-SQL. *CoRR*, abs/2507.22478, 2025. doi: 10.48550/ARXIV.2507.22478. URL <https://doi.org/10.48550/arXiv.2507.22478>. arXiv: 2507.22478.
- Shi, Y., Wang, S., Wan, C., and Gu, X. From Code to Correctness: Closing the Last Mile of Code Generation with Hierarchical Debugging. *CoRR*, abs/2410.01215, 2024. doi: 10.48550/ARXIV.2410.01215. URL <https://doi.org/10.48550/arXiv.2410.01215>. arXiv: 2410.01215.
- Shojaee, P., Jain, A., Tipirneni, S., and Reddy, C. K. Execution-based Code Generation using Deep Reinforcement Learning. *Trans. Mach. Learn. Res.*, 2023, 2023. URL <https://openreview.net/forum?id=0XBuaxqEcG>.
- Song, D., Zhou, Z., and Wang, Z. An Empirical Study of Code Generation Errors made by Large Language Models. 2023.
- Song, Y., Lothritz, C., Tang, X., Bissyandé, T. F., and Klein, J. Revisiting Code Similarity Evaluation with Abstract Syntax Tree Edit Distance. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Short Papers, Bangkok, Thailand, August 11-16, 2024*, pp. 38–46. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-SHORT.3. URL <https://doi.org/10.18653/v1/2024.acl-short.3>.
- Wang, X., Wang, Y., Wan, Y., Mi, F., Li, Y., Zhou, P., Liu, J., Wu, H., Jiang, X., and Liu, Q. Compilable Neural Code Generation with Compiler Feedback. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 9–19. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.2. URL <https://doi.org/10.18653/v1/2022.findings-acl.2>.
- Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL <https://doi.org/10.48550/arXiv.2412.15115>. arXiv: 2412.15115.
- Ye, Y., Zhang, T., Jiang, W., and Huang, H. Process-Supervised Reinforcement Learning for Code Generation. *CoRR*, abs/2502.01715, 2025. doi: 10.48550/ARXIV.2502.01715. URL <https://doi.org/10.48550/arXiv.2502.01715>. arXiv: 2502.01715.
- Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., Zhang, Z., and Radev, D. R. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 3911–3921. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1425. URL <https://doi.org/10.18653/v1/d18-1425>.
- Zhang, J., Zhang, J., Li, Y., Pi, R., Pan, R., Liu, R., Zheng, Z., and Zhang, T. Bridge-Coder: Transferring Model Capabilities from High-Resource to Low-Resource Programming Language. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 10865–10882. Association for Computational Linguistics, 2025a. URL <https://aclanthology.org/2025.findings-acl.567/>.

Zhang, K., Li, Z., Li, J., Li, G., and Jin, Z. Self-Edit: Fault-Aware Code Editor for Code Generation. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 769–787. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.45. URL <https://doi.org/10.18653/v1/2023.acl-long.45>.

Zhang, Z., Wang, C., Wang, Y., Shi, E., Ma, Y., Zhong, W., Chen, J., Mao, M., and Zheng, Z. LLM Hallucinations in Practical Code Generation: Phenomena, Mechanism, and Mitigation. *Proc. ACM Softw. Eng.*, 2(ISSTA):481–503, 2025b. doi: 10.1145/3728894. URL <https://doi.org/10.1145/3728894>.

Zhong, R., Yu, T., and Klein, D. Semantic evaluation for text-to-sql with distilled test suite. In *The 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2020.

Zhou, S., Alon, U., Agarwal, S., and Neubig, G. Code-BERTScore: Evaluating Code Generation with Pre-trained Models of Code. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 13921–13937. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.859. URL <https://doi.org/10.18653/v1/2023.emnlp-main.859>.

A ANSIBLE DATASET STATISTICS

Category	Count
Repositories Scraped from Old Ansible Galaxy	34057
Repositories Scraped from New Ansible Galaxy	3212
Total Repositories Scraped	37269
Total Playbooks Extracted	74497
Playbooks with Undefined Variables	5906
Playbooks with Duplicate Contents	35033
Playbooks Successfully Parsed	33558

Table 5. Summary of Ansible repository and playbook collection statistics.

B PROMPT TEMPLATES

B.1 Code Generation Prompts

Ansible Generation Prompt

You are an expert in Ansible. The user will give you a task description and ask you to generate an Ansible playbook to complete the given task. You only need to output the content of the playbook. DO NOT use any shell commands (ansible.builtin.shell, ansible.builtin.command, etc.) in the playbook.

Task: {task}

Answer: “yaml

Bash Generation Prompt

You are an expert in Bash. The user will give you a task description and ask you to generate a bash command to complete the given task. You only need to output the content of the command.

Task: {task}

Answer: “bash

SQL Generation Prompt

You are an expert in SQL. The user will give you a task description and ask you to generate a SQL command to complete the given task. You only need to output the content of the command.

Task: {task}

Answer: “sql

B.2 Program Repair Prompts

Ansible Program Repair Prompt

You are an expert in Ansible. You are asked to fix a possibly incorrect Ansible playbook. You will be provided with the playbook to fix, the user input, and feedback from an interpreter that lists all syntactic errors in the playbook. Your goal is to fix the syntactic errors in the playbook (if any) while following the user's instruction. You only need to output the content of the modified playbook.

User query: {query}

Original playbook:
{output}

Interpreter feedback:
{feedback}

Answer: “yaml

SQL Program Repair Prompt

You are an expert in SQL. You are asked to fix a possibly incorrect SQL command. You will be provided with the command to fix, the user input, and feedback from an interpreter that lists all syntactic errors in the command. Your goal is to fix the syntactic errors in the command (if any) while following the user's instruction. You only need to output the content of the modified command.

User query: {query}

Original command:
{output}

Interpreter feedback:
{feedback}

Answer: “sql

B.3 In-context Learning Prompts

Ansible In-context Learning Prompt

You are an expert in Ansible. The user will give you a task description and ask you to generate an Ansible playbook to complete the given task. You only need to output the content of the playbook. DO NOT use any shell commands (ansible.builtin.shell, ansible.builtin.command, etc.) in the playbook. The following are some example input queries and corresponding Ansible playbooks for your reference:

{examples}
Task: {task}
Answer: “yaml

User query: {query}

Original command:
{output}

Interpreter feedback:
{feedback}

Answer: “bash

Bash In-context Learning Prompt

You are an expert in Bash. The user will give you a task description and ask you to generate a bash command to complete the given task. You only need to output the content of the command.

The following are some example input queries and corresponding Bash commands for your reference:
{examples}
Task: {task}
Answer: “bash

SQL In-context Learning Prompt

You are an expert in SQL. The user will give you a task description and ask you to generate a SQL command to complete the given task. You only need to output the content of the command.

The following are some example input queries and corresponding SQL commands for your reference:

{examples}

Task: {task}

Answer: ““sql

B.4 Ansible Dataset Query Generation Prompt

Natural Language Query Generation Prompt

You are an expert in Ansible. You are asked to write a user prompt for the given Ansible playbook that can be used to generate the playbook. Instead of explicitly describing the functionality of the playbook, the prompt should tell what the user wants to accomplish through the playbook. Write the prompt as short as you can, and start the prompt with: Generate an Ansible playbook that ...