
Divide-and-Conquer Meets Consensus: Unleashing the Power of Functions in Code Generation

Jingchang Chen*

Harbin Institute of Technology
jcchen@ir.hit.edu.cn

Hongxuan Tang*

Harbin Institute of Technology
jeffswt@outlook.com

Zheng Chu

Harbin Institute of Technology
zchu@ir.hit.edu.cn

Qianglong Chen[†]

Zhejiang University
chenqianglong.ai@gmail.com

Zekun Wang

Harbin Institute of Technology
zkwang@ir.hit.edu.cn

Ming Liu[†]

Harbin Institute of Technology
mliu@ir.hit.edu.cn

Bing Qin

Harbin Institute of Technology
qbinqin@ir.hit.edu.cn

Abstract

Despite recent progress made by large language models in code generation, they still struggle with programs that meet complex requirements. Recent work utilizes plan-and-solve decomposition to decrease the complexity and leverage self-tests to refine the generated program. Yet, planning deep-inside requirements in advance can be challenging, and the tests need to be accurate to accomplish self-improvement. To this end, we propose FUNCODER, a code generation framework incorporating the divide-and-conquer strategy with functional consensus. Specifically, FUNCODER recursively branches off sub-functions as smaller goals during code generation, represented by a tree hierarchy. These sub-functions are then composited to attain more complex objectives. Additionally, we designate functions via a consensus formed by identifying similarities in program behavior, mitigating error propagation. FUNCODER outperforms state-of-the-art methods by +9.8% on average in HumanEval, MBPP, xCodeEval and MATH with GPT-3.5 and GPT-4. Moreover, our method demonstrates superiority on smaller models: With FUNCODER, StableCode_{3b} surpasses GPT-3.5 by +18.6% and achieves 97.7% of GPT-4's performance on HumanEval. Further analysis reveals that our proposed dynamic function decomposition is capable of handling complex requirements, and the functional consensus prevails over self-testing in correctness evaluation.

1 Introduction

Over the past few years, large language models have been observed to attain significant advancements in coding capabilities (OpenAI, 2023; Touvron et al., 2023). Meanwhile, models designed specifically for coding tasks have also been introduced (Rozière et al., 2023; Lozhkov et al., 2024; Pinnaparaju et al., 2024). Although LLMs can proficiently generate simple code snippets, they suffer from a decline in performance as code requirements become complicated.

Numerous efforts have been made to tackle this complexity. The two-stage methods (Jiang et al., 2023; Zelikman et al., 2023) employ the plan-and-solve strategy, which first generates a draft outline

*Equal contribution.

[†]Corresponding Authors: Ming Liu, Qianglong Chen.

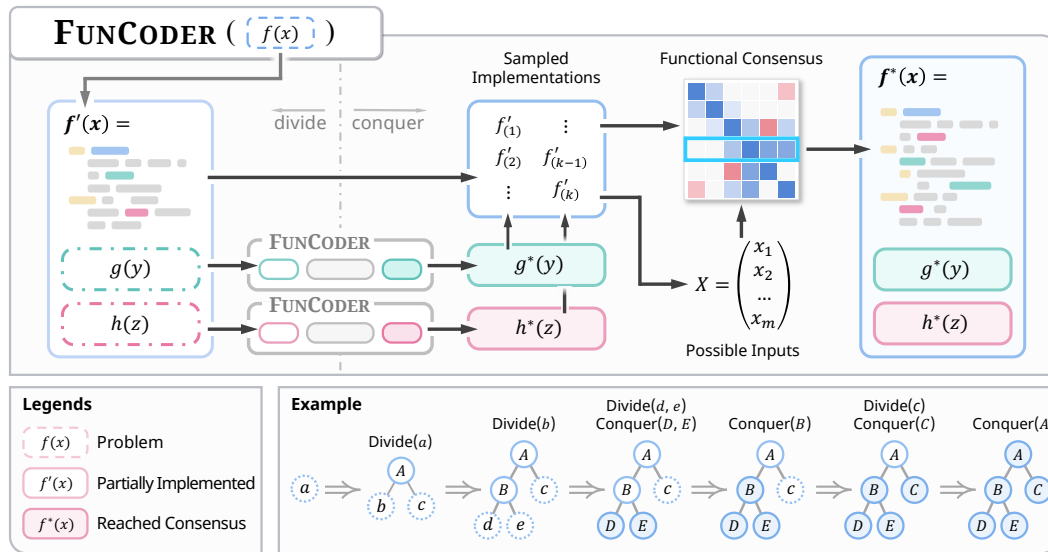


Figure 1: A flowgraph illustrates FUNCODER. FUNCODER branches off new functions to have sub-goals tackled iteratively (left), re-composites sub-functions, and selects the best using functional consensus (right). Bottom-right figure shows how FUNCODER writes functions at hierarchy-level.

for the complex task and uses it as guidance for implementing the code in the second stage. Multi-agent development frameworks (Hong et al., 2024; Qian et al., 2023) mimic real-world software development workflows, assign different roles to LLMs and collaborate to solve a complex goal. Self-improvement (Shinn et al., 2023; Chen et al., 2024), on the other hand, refines the program in accordance with execution feedback from self-generated unit tests.

Despite fruitful efforts made by the previous methods in dealing with complex problems, certain challenges still remain unsolved: (1) Two-stage approaches need to design a complete plan at the beginning and lack the ability to adjust the top-level design during implementation, leading to sub-optimal decomposition. (2) Multi-agent collaboration frameworks are cumbersome and rely heavily on LLM capabilities, making them difficult to generalize to smaller open-source models. (3) Code refinement through self-tests depends on the correctness of generated unit-tests. Our preliminary study (§3.1.3) finds that models generate unreliable self-tests in abundance. These incorrect tests may mislead self-improvement and, at worse, exacerbate program errors.

To address these issues, we propose FUNCODER, a code generation framework utilizing a divide-and-conquer strategy and a novel functional consensus mechanism on functions to decompose complex problems. Starting from the main problem, FUNCODER introduces new functions to cope with certain sub-problems. The new functions will be decomposed recursively, eventually forming a tree of functions. FUNCODER then combines functions bottom-up to achieve increasingly complicated objectives. By dividing-and-conquering tasks into simpler sub-functions, complexity can be gradually reduced. However, errors in sub-functions may propagate to the whole program, thereby damaging overall reliability. We propose functional consensus that samples multiple functions and selects the one demonstrating consensus, measured by the aggregated similarity among candidates. By reaching a consensus, we reduce the discrepancies in code behavior and thus alleviate cascading errors.

We conduct extensive experiments on code generation benchmarks (Chen et al., 2021; Austin et al., 2021; Khan et al., 2023) with GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023), outperforming state-of-the-art methods by +9.8% on average. Experiments are further carried out on the mathematical competition benchmark, MATH (Hendrycks et al., 2021b), achieving a +6.0 improvement with GPT-4, indicating that FUNCODER can also generalize to complex reasoning. Our method is observed to be equally effective on open-source models (Meta AI, 2024; Mistral AI, 2024; Pinnaparaju et al., 2024; Rozière et al., 2023; Lozhkov et al., 2024), with an average gain over baseline of +31.5% on HumanEval and +47.7% on MATH. Additional analysis also shows the advantage of both divide-and-conquer and functional consensus. Our code is made openly available at <https://github.com/cometeme/funcoder>.

Algorithm 1 FUNCODER procedure**Require:** Entry func, $f_{\text{root}} = \{h_{\text{root}}, d_{\text{root}}, \phi\}$ **Require:** Large language model, LLM

```

1: function FUNCODER( $f_{\text{cur}}$ )
2:   — Divide —
3:    $f'_{\text{cur}}, \{f_i\} \leftarrow \text{EXTRACT}(\text{LLM}(f_{\text{cur}}))$ 
4:   for  $f_i \in \{f_i\}$  do
5:     if  $b_i$  is NOTIMPLEMENTED then
6:        $f_i^* \leftarrow \text{FUNCODER}(f_i) \triangleright$  recursion
7:     end if
8:      $\text{ADDCHILD}(f_{\text{cur}}, f_i^*)$ 
9:   end for
10:  — Conquer —
11:   $F_{\text{cur}} \leftarrow \text{SAMPLE}(\text{LLM}(f'_{\text{cur}}, \text{CHILD}(f_{\text{cur}})))$ 
12:   $f_{\text{cur}}^* \leftarrow \text{FUNCONSENSUS}(F_{\text{cur}})$ 
13:  return  $f_{\text{cur}}^*$ 
14: end function
15: return FUNCODER( $f_{\text{root}}$ )  $\triangleright$  starts from root

```

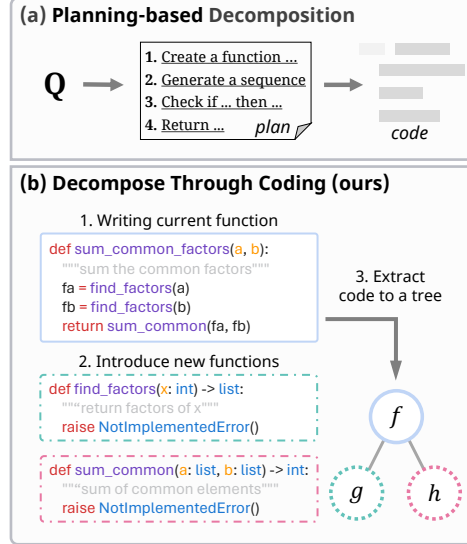


Figure 2: Left: Algorithm for FUNCODER, explained in detail in Appendix A.6. Right: Comparison between decomposition by planning and our approach. FUNCODER introduces new functions to describe sub-goals solely with code, achieving a more natural way of requirement decomposition.

2 FUNCODER: Divide-and-Conquer Meets Consensus

2.1 Divide-and-Conquer for Iterative Programming

A function is defined as a relation between a set of inputs and outputs where each input is assigned exactly one output (Halmos, 1998), denoted as $y = f(x)$. In computer programming, a function is identified by its header h_f with its body b_f , and is commonly accompanied by a documentation d_f to improve readability. Functions can be invoked from other procedures, allowing for the decomposition of large and complicated requirements into smaller structures that exhibit high comprehensibility and quality (Dahl et al., 1972). Generally, human programmers tend to decompose tasks into clearly defined sub-functions and then implement them recursively, making functions eligible for re-usage, taking advantage of the *divide-and-conquer* principle. Inspired by this, FUNCODER recursively *divides* the requirement and *conquers* functions to formulate a sophisticated solution, unleashing the potential of LLMs in code generation.

Divide is a top-down process that iteratively breaks down problems. Given a code generation problem, the process begins from the entry function f_{root} . We instruct the model to introduce new functions $f_i \in \text{CHILD}(f_{\text{cur}})$ that solve certain sub-goals while writing the current f_{cur} . To reduce the complexity involved in each generation, we only require the headers h_{f_i} and documentation d_{f_i} of new functions to be generated, while their implementations b_{f_i} can be postponed. After completing the current function, the model starts to address those unimplemented sub-functions and complete b_{f_i} into f'_i . This process stops when the model deems functions too simple to be further divided, finally forming a dependency tree $T = \text{TREE}(f_{\text{root}}, \text{CHILD}(f_{\text{root}}))$. The *divide* process is similar to a search starting from the entry function, gradually involving new sub-functions while writing the current, and implementing them recursively. We guide the entire process through a depth-first search.

Conquer is a process of achieving complex objectives through aggregating smaller functions. We notice that child functions are not yet implemented during the top-down process of writing parent functions. As a result, these parent functions may not be able to effectively utilize the child functions, or misuse them at worst. FUNCODER deals with this issue by re-generating functions in inverse topological order on the dependency tree T - starting from leaves, complex goals are handled by compositing solved children as $f_{\text{cur}}^* \leftarrow \mathcal{F}(f'_{\text{cur}}, \{f_1^*, f_2^*, \dots\}) \mid f_i^* \in \text{CHILD}(f_{\text{cur}})$.

Divide and *conquer* naturally achieve both decomposition and composition during code generation. Unlike two-stage and agent-based methods, our approach dynamically introduces new functions

along the process, making it less burdensome than producing a complete plan at the very beginning. Moreover, while planning or agents require chat capabilities, FUNCODER represents sub-tasks through functions (Figure 2), making it more applicable to specialized code generation models.

2.2 Functionality Similarity as a Consensus

The decomposition of complex tasks benefits from solving easier sub-goals, but might introduce the risks of cascading errors, which refers to errors in sub-functions that lead to errors in ancestral functions. To mitigate this, we introduce Functional Consensus which aims at reducing inconsistencies in program behavior. This is achieved by sampling multiple functions and selecting the one that exhibits consensus, as measured by the aggregated similarity of functionality between candidates, thus abating outlier functionalities.

Functionality Similarity A program specifies its functionality (or behavior) through the control flow and logic defined by its code semantics. However, comparing the functionalities between two programs based on their semantics is somewhat challenging. By decomposing the requirement into functions, FUNCODER is able to view the function behavior as a black box that maps arguments into return values. Considering two functions f and g with the same input domain $D(f) = D(g)$, we define the similarity between them $\text{sim}(f, g)$ as the identicalness of outputs when given the same input values.

$$\text{sim}(f, g) = \int_{x \in D(f)} \frac{\mathbb{1}[f(x) = g(x)]}{|D(f)|} \approx \sum_{x \in X | X \sim D(f)} \frac{\mathbb{1}[f(x) = g(x)]}{|X|} \quad (1)$$

The similarity becomes 1 if and only if two functions output consistent values for all inputs: $\forall x \in D(f) : f(x) = g(x) \Leftrightarrow \text{sim}(f, g) = 1$. We notice that the input domain $D(f)$ is unbounded in most cases, making its measurement barely feasible in practice. Thus, we approximate it by sampling a subset of possible inputs $X \sim D(f)$ with an LLM.

Consensus is reached by selecting the candidate f^* holding maximal similarity with others after sampling multiple function implementations $F = \{f_{(i)}\}$ for the same requirements.

$$f^* = \text{FUNCONSENSUS}(F) = \arg \max_{f_{(i)} \in F} \sum_{f_{(j)} \in F \setminus \{f_{(i)}\}} \text{sim}(f_{(i)}, f_{(j)}) \quad (2)$$

By introducing functional consensus, FUNCODER produces functions that are more consistent and common in functionality, while omitting abnormal samples. The process is applied to not just the final program, but also to every sub-tree during the bottom-up *conquering* stage, resulting in step-by-step, thorough verification from the most fundamental functions all the way up to the whole program.

2.3 FUNCODER is a Function Coder

We design FUNCODER as a procedure that takes a problem in the form of a function signature $f(x)$, and produces a final solution $f^*(x)$, as exemplified in Figure 1. Given a problem $f(x)$, FUNCODER partially implements the function as $f'(x)$ referring to unimplemented sub-functions $g(y)$ and $h(z)$. These sub-functions are then fed into FUNCODER to be recursively coped with. We then sample k implementations $f'_{(i)}(x)$ based on solved children $g^*(y)$ and $h^*(z)$. Functional consensus is calculated by evaluating candidates on possible inputs. The function sharing maximal behavioral similarity is combined with solved children to formulate the final solution.

3 Experiments

We conduct experiments on competition-level code generation and mathematical reasoning benchmarks with state-of-the-art LLMs, which are covered in section §3.1 and §3.2, respectively. In addition to GPT models (Ouyang et al., 2022; OpenAI, 2023), we also conduct experiments with community models like Llama3_{8b} (Meta AI, 2024), StableCode_{3b} (Pinnaparaju et al., 2024), and CodeLlama_{34b} (Rozière et al., 2023). We use the *instruct* variant of these models and inference on a single A100-80G under BF16 precision with vLLM (Kwon et al., 2023).

Table 1: Experiment results on code generation benchmarks. We report Pass@1 as evaluate metric. Results from the original paper are underlined, and the best results are bold.

Model	Method	HumanEval		MBPP		xCodeEval				
		Pass@1	$\Delta \uparrow$	Pass@1	$\Delta \uparrow$	Easy	Mid	Hard	Expert	All
GPT-3.5	Standard	68.3	-	72.0	-	44.4	15.2	4.6	0.0	20.2
	CodeT	81.1	+12.8	76.0	+4.0	50.6	16.1	8.0	0.0	23.2
	Reflexion	69.5	+1.2	72.5	+0.5	44.4	17.0	5.7	0.0	20.6
	LDB	<u>82.9</u>	+14.6	<u>76.0</u>	+4.0	-	-	-	-	-
	FUNCODER	85.4	+17.1	78.5	+6.5	62.4	29.5	11.6	0.0	31.4
GPT-4	Standard	82.9	-	73.5	-	68.5	39.3	19.5	1.7	37.4
	Parsel	<u>85.0</u>	+2.1	-	-	-	-	-	-	-
	CodeT	90.9	+8.0	77.0	+3.5	76.4	51.8	21.8	3.4	44.0
	Reflexion	<u>91.0</u>	+8.1	<u>77.1</u>	+3.6	71.3	41.1	19.5	2.5	38.6
	MetaGPT	<u>85.9</u>	+3.0	-	-	-	-	-	-	-
	FUNCODER	94.5	+11.6	79.5	+6.0	83.1	58.0	26.4	3.4	48.6
Llama3 _{8b}	Standard	61.6	-	60.5	-	9.0	1.8	0.0	0.0	3.6
	CodeT	68.9	+7.3	61.5	+1.0	12.4	0.0	0.0	0.0	4.4
	FUNCODER	79.7	+18.1	62.5	+2.0	22.0	0.9	0.0	0.0	8.0
StableCode _{3b}	Standard	61.0	-	51.5	-	7.3	0.9	0.0	0.0	2.8
	CodeT	75.0	+14.0	57.5	+6.0	11.2	1.8	0.0	0.0	4.6
	FUNCODER	81.0	+20.0	63.5	+12.0	13.5	4.5	1.1	0.0	6.2
CodeLlama _{34b}	Standard	43.9	-	53.5	-	2.3	0.0	0.0	0.0	0.8
	CodeT	55.5	+11.6	56.5	+3.0	10.1	0.0	0.0	0.0	3.6
	FUNCODER	66.5	+22.6	58.5	+5.0	10.2	0.0	0.0	0.0	3.6

3.1 Code Generation

We choose three benchmarks for code generation evaluation: (a) HumanEval (Chen et al., 2021) includes entry-level coding questions; (b) MBPP (Austin et al., 2021) contains questions of standard library invocation and programming basics; and (c) xCodeEval (Khan et al., 2023) consists of algorithmic challenges sourced from the competitive programming platform CodeForces.

3.1.1 Experiment Setup

Benchmarks We adopt the full test set (164 problems) for HumanEval, and sample 200 for MBPP and 500 for xCodeEval, respectively. Following EbTech (2024), we split the xCodeEval into 4 subsets based on problem difficulty: Easy (≤ 1200), Mid (1200-1599), Hard (1600-1999) and Expert (≥ 2000). The evaluation metric for code generation is Pass@1 unless specified.

Baselines We compare FUNCODER with standard prompting (Brown et al., 2020), two-stage decomposition method Parsel (Zelikman et al., 2023), self-testing method CodeT (Chen et al., 2023a), self-improvement methods Reflexion and LDB (Shinn et al., 2023; Zhong et al., 2024), and multi-agent developing framework MetaGPT (Hong et al., 2024). We implement Standard prompting with a 1-shot demonstration. CodeT samples 11 solutions with standard prompting and evaluates them on model-generated tests. The results for Reflexion are reproduced from the original code.

Implementation Details FUNCODER uses a 2-shot prompt in the divide stage and 1-shot for conquering sub-functions. The number of sampled implementations in the functional consensus is set to 11 for code generation tasks. For further implementation details, please refer to Appendix A.1.

3.1.2 Results

Table 1 shows the code generation performance on advanced proprietary models, GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023). For basic programming questions, HumanEval and MBPP, FUNCODER surpass previous SOTA methods by +3.3% in Pass@1 and reduce the error rate by 18.6%. Furthermore, FUNCODER demonstrates a substantial improvement on competition-level problems, outperforming others by 10.4% in GPT-4 and 35.3% with GPT-3.5. We observe that FUNCODER can

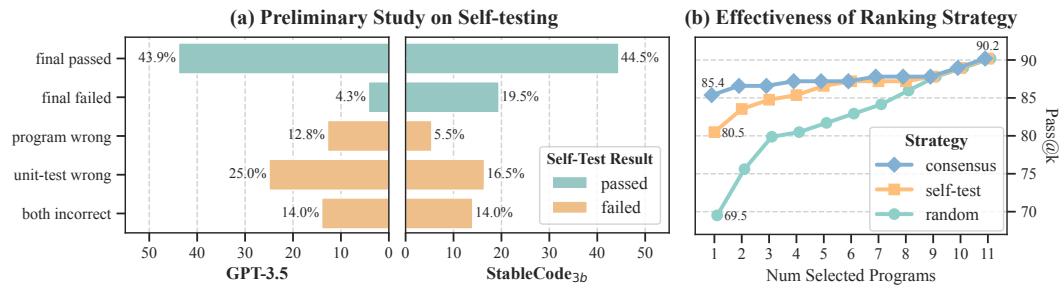


Figure 3: (a) Preliminary study on self-testing, the programs are evaluated using unit-tests generated by LLMs. (b) The effectiveness of different ranking strategies. We compute the Pass@k over top-k programs ranked by functional consensus, self-test, and random on 11 candidates. (higher is better)

enhance LLM’s capability of solving more complex programming tasks, with an average accuracy improvement of 82.3% over the baseline on the *Mid* and *Hard* subsets of xCodeEval. *Expert* level programs, however, still remain a colossal challenge for even the most cutting-edge LLMs.

Evaluation is also performed over community LLMs, Llama3 (Meta AI, 2024), StableCode (Pinaparaju et al., 2024) and CodeLlama (Rozière et al., 2023) with results in Table 1. FUNCODER consistently boosts the performance of smaller models in code generation, demonstrating notable improvements compared to standard prompting on HumanEval, which gained +29.4% on Llama3, +32.8% on StableCode, and even +51.5% on CodeLlama, outperforming that from the previous best method CodeT. We also supplement results on GPT-4o mini, Codestral and StarCoder2 in Table 11. Experiment results demonstrate that our method archives state-of-the-art performance on various models, ranging from basic programming to competition contests.

3.1.3 Analysis

FUNCODER Democratize to Smaller LLMs Limited by the LLM capabilities, the application of self-improvement or multi-agent methods on smaller models is without ease. By keeping decomposition and composition within the code generation process, our approach exhibits better generalization. As shown in Table 1, with FUNCODER, StableCode_{3b} achieves around 118.6% relative performance to standard GPT-3.5, and also aligns closely with GPT-4 by about 97.7% on HumanEval.

Preliminary Study on Self-Testing Method We conduct a preliminary study targeting the self-testing method on HumanEval, results are shown in Figure 3.a with further details in Appendix A.5. We first verify whether model-generated programs can also pass model-generated self-tests: (a) If a program passes self-tests, most from GPT-3.5 would also work on system tests, as much as $19.5\%/64\% \approx 30.5\%$ programs from StableCode are rejected, indicating that smaller models like StableCode may not effectively self-test and detect program errors on its own. (b) In the event of failed self-tests, a large portion of failures are attributed to issues in self-tests instead of the programs, on both GPT-3.5 and StableCode. These phenomena indicate that self-testing methods have limitations in generating correct and reliable unit tests. As a result, we design functional consensus to not require any assertion, but perform *mutual verification* between solutions instead, as opposed to self-testing.

Effectiveness of Functional Consensus Functional consensus or self-testing may be viewed as ranking algorithms for selecting functions. To measure ranking effectiveness, we conduct an analysis on HumanEval with GPT-3.5. For each problem, 11 candidates are ranked with 3 strategies: consensus, self-test, and random shuffle (as a baseline). Effectiveness is measured via Pass@k, i.e. if any of the top-k ranked programs pass the system test. Figure 3.b shows that functional consensus achieves 94.7% upper bound (Pass@11) performance by selecting a *single* function (Pass@1), and is close to that of self-test on Pass@4. This clearly demonstrates that functional consensus can effectively evaluate correctness and pick the most promising implementation on the first attempt.

Ablation and Token Usage To analyze the impact of dividing, conquering, and functional consensus in FUNCODER, we carry out an ablation study with different settings. Studies that replace consensus with self-testing, or with AlphaCode-like (Li et al., 2022) clustering, are also included. The ablation is constructed on HumanEval with GPT-3.5, as shown in Table 2. Note that to generate every program FUNCODER costs only $O(kN)$ tokens, where k is the number of sampled candidates,

Table 2: Ablation study of FUNCODER on HumanEval with GPT-3.5. The setting in our main experiment is highlighted in bold. Tokens are calculated as the sum of prompts and completions.

Setting	Divide	Conquer	Ranking	Pass@1	Avg. Tokens
Standard	✗	✗	✗	68.3	886.7
One-pass	✓	✗	✗	72.6 (+4.3)	1233.7
Two-pass	✓	✓	✗	78.7 (+10.4)	3343.2
Two-pass + ST@11	✓	✓	Self-Test@11	80.5 (+12.2)	5408.3
Two-pass + CL@11	✓	✓	Clustering@11	75.0 (+6.7)	5070.7
FUNCODER@5	✓	✓	Consensus@5	83.5 (+15.2)	4040.9
FUNCODER@11	✓	✓	Consensus@11	85.4 (+17.1)	5402.0

and N is the token length of the final program. This is further exemplified and explained in §A.7. We observe that function decomposition and re-composition deliver cumulative performance improvements. Functional consistency is also shown to prevail over self-testing. Putting them all together, FUNCODER received a +17.1 improvement with just $5.09\times$ more tokens over baseline. Compared to previous SOTA LDB ($\approx 23K$ tokens), we are able to gain +2.5 in performance with 76.5% token usage reduction.

3.2 Mathematical Reasoning

Code can be viewed as a tool for augmenting the reasoning capabilities of LLMs (Chen et al., 2023b). Alternative to text-based reasoning like Chain-of-Thought (Wei et al., 2022), programs can offer unique advantages in terms of iteration and calculations. To test the generalizability of FUNCODER beyond algorithm challenges, we conduct an experiment on MATH (Hendrycks et al., 2021b), a competition-level mathematical reasoning benchmark.

3.2.1 Experiment Setup

Benchmark The experiment is conducted on a subset of the MATH test set, including 500 randomly sampled problems that can be classified into 7 disjoint subjects or 5 difficulty levels. It can be noticed that labels in MATH are formatted in $\text{\LaTeX}$, rendering exact-match verdicts impractical. We, therefore, follow previous work (Zhang et al., 2024) and adopt GPT-4 to determine the correspondence between predictions and labels, with further details provided in Appendix A.4.

Baselines We compare FUNCODER with the text-based baselines: Standard Prompting and Chain-of-Thought (Wei et al., 2022), and program-aided baselines: Program-of-Thought (Chen et al., 2023b), Self-Refine (Madaan et al., 2023), Cumulative Reasoning (Zhang et al., 2024). The results of Cumulative reasoning are reported in the original paper. Standard prompting and chain-of-thought reasoning use 7-shot demonstrations constructed from the train set. Program-of-Thought and Self-Refine prompt the model with 1-shot demonstration to generate a `solution()` function that solves the problem. Additionally, self-refine iteratively refines programs based on runtime feedback. All baseline methods are run with self-consistency (Wang et al., 2023) at 5.

Implementation Details FUNCODER adopts a program-aided reasoning setting that writes a `solution()` function and obtains the final prediction by running this program. The number of sampled implementations $|F|$ in functional consensus is set to 5 to match baseline methods.

3.2.2 Results

The experimental results on MATH are shown in Table 3. It shows that program-aided reasoning generally outperforms text-based reasoning. With GPT-4 as the backbone, FUNCODER outperforms the strongest baseline Cumulative Reasoning (Zhang et al., 2024) by (6.0 / 8.3%) and surpasses the vanilla program-aided baseline PoT (Chen et al., 2023b) by (10.0 / 14.7%). When using GPT-3.5-turbo as the backbone, FUNCODER exceeds the strongest baseline by (6.2 / 11.1%) and outperforms PoT by as much as (13.0 / 31.7%), which indicates that our approach has a strong advantage over both text-based reasoning and other program-aided reasoning methods.

Table 3: Experimental results on MATH, a competition-level mathematical reasoning benchmark. Best results are in bold. Text-based reasoning methods are denoted with $\dagger$, while others use program-aided reasoning. We report both overall results and results in seven subjects: *Prealgebra*, *Algebra*, *Number Theory*, *Counting & Probability*, *Geometry*, *Intermediate Algebra*, and *Precalculus*.

Model	Method	<i>Prealg.</i>	<i>Alg.</i>	<i>NT</i>	<i>Prob.</i>	<i>Geo.</i>	<i>InterAlg.</i>	<i>Precalc.</i>	Overall
GPT-3.5	Standard $\dagger$	62.2	37.4	20.0	29.8	31.0	24.4	21.8	34.6
	CoT $\dagger$	59.8	51.1	28.9	29.8	28.6	26.7	30.9	40.0
	PoT	68.3	50.4	33.3	48.9	21.4	18.2	29.1	41.0
	Self-Refine	74.4	49.6	48.9	57.4	28.6	35.6	36.4	48.6
	FUNCODER	76.8	61.2	55.6	59.6	34.1	36.0	41.8	54.0
GPT-4	Standard $\dagger$	81.7	82.7	71.1	72.3	59.5	46.7	47.3	68.2
	CoT $\dagger$	84.1	87.1	62.2	68.1	45.2	48.9	54.5	68.6
	PoT	79.3	80.6	75.6	72.3	50.0	47.8	58.2	68.2
	Self-Refine	82.9	82.0	77.8	76.6	54.8	55.6	63.6	72.2
	CR	86.6	86.3	88.7	71.1	53.7	51.5	51.8	72.2
	FUNCODER	89.0	92.8	82.2	83.0	59.5	63.3	56.4	78.2
Llama3 _{8b}	CoT $\dagger$	56.1	47.5	31.1	34.0	40.5	14.4	38.2	38.6
	PoT	67.1	32.4	24.4	34.0	16.7	21.1	18.2	32.6
	FUNCODER	67.9	45.7	51.1	53.2	19.0	37.8	30.9	45.0
StableCode _{3b}	PoT	20.7	14.4	17.8	25.5	4.8	8.9	9.1	14.4
	FUNCODER	46.3	30.2	20.0	29.8	4.8	20.0	18.2	26.6
CodeLlama _{34b}	PoT	35.5	26.1	15.0	16.7	0.0	5.5	33.3	15.2
	FUNCODER	44.8	46.1	37.8	34.1	13.6	24.6	37.5	24.4

On open-source models, FUNCODER with Llama3 outperforms PoT by (12.4 / 38.0%). It has even reached competitive performance against the state-of-the-art method based on GPT-3.5 (45.0 v.s. 48.6). When employing StableCode and CodeLLaMA as the backbone, our approach achieves significant improvements by (12.2 / 84.7%) and (9.2 / 60.5%), respectively. This improvement demonstrates that our approach can significantly boost smaller LLMs, democratizing the complex reasoning capabilities of open-source LLMs through programming.

3.2.3 Analysis

FUNCODER Can Handle Harder Questions

Figure 4 compares between CoT, PoT, and FUNCODER across varying difficulty levels. It illustrates that CoT performs comparatively well on the easiest questions, but suffers from a steep decline in performance as difficulty increases. This suggests that text-based reasoning is inadequate for tackling challenging mathematical reasoning problems. The same situation is also observed in PoT. In contrast, our method consistently demonstrates high performance even on challenging problems, particularly excelling on level 5 difficulty with nearly double the performance compared to PoT and CoT. This reflects that our method, with divide-and-conquer applied, can effectively cope with complex problems.

Decomposed Functions are Domain-Specific We hypothesize that questions from the same subject require similar knowledge reserves, which should be reflected in the functionality of the sub-functions. To verify this hypothesis, we statisticize the common sub-functions of FUNCODER in each MATH subject, as shown in Table 4. It is apparent that different subjects require different abilities, each with its own set of sub-functions closely associated with the domain knowledge. In addition, these common sub-functions are fundamentally basic and straightforward. As exemplified in Appendix B.2, our method is able to leverage and combine these basic sub-functions to achieve more complex goals, thereby reducing the complexity of reasoning and enhancing performance.

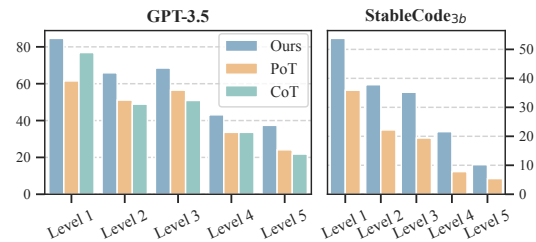


Figure 4: Average accuracy in each level with the chat model (GPT-3.5) and the code model (StableCode_{3b}) on the MATH benchmark.

Table 4: Top-3 most commonly used functions in each subject of MATH, listed in descending order.

Subject	Functions
Prealgebra	is_prime / factorial / gcd
Algebra	find_roots / is_perfect_square / find_domain
Number Theory	get_divisors / mod_inverse / gcd
Counting & Probability	factorial / combinations / binomial_coefficient
Geometry	distance / simplify_fraction / calculate_triangle_area
Intermediate Algebra	find_roots / evaluate_polynomial / lagrange_interpolation
Precalculus	cross_product / fraction_from_angle / dot

4 Related Work

Large Language Model for Code Code pre-training has received widespread attention, with early models based on small language models (SLM) (Feng et al., 2020; Lu et al., 2021; Wang et al., 2021). In recent years, with the development of large-scale pre-training techniques, code LLM has emerged, showing remarkable performance in downstream code tasks (Chen et al., 2021; Nijkamp et al., 2023; Li et al., 2022; Rozière et al., 2023; Li et al., 2023b; Guo et al., 2024). Tasks between code and natural language (NL) can be generally divided into three major categories: NL2Code tasks such as code generation (Austin et al., 2021; Chen et al., 2021; Hendrycks et al., 2021a; Khan et al., 2023) and code search (Husain et al., 2019); Code2Code tasks including code completion (Lu et al., 2021; Zhang et al., 2023; Liu et al., 2024), code translation (Ahmad et al., 2023; Zhu et al., 2022; Yan et al., 2023), and test generation (Siddiq et al., 2023; Schäfer et al., 2024); Code2NL tasks like code summarization (Jin et al., 2023). This paper focuses on code generation tasks, ranging from basic to competition level.

Code Refinement and Self-Testing Code doesn't always run as expected; it could contain syntax errors, dead loops, or bugs. It's essential to debug and refine the code to ensure better quality. CodeT (Chen et al., 2023a) generates unit-tests to score the implementation. AlphaCode (Li et al., 2022) clusters programs based on whether generated program outputs were identical or not. Self-improvement methods (Madaan et al., 2023; Shinn et al., 2023; Chen et al., 2024; Zhong et al., 2024) design closed-loop procedures that repeatedly refine the code based on the feedback. Like real-life software development processes, multi-agent frameworks (Hong et al., 2024; Qian et al., 2023) construct specific LLM roles, *Tester* or *QA* to generate tests. These studies adopt a shared paradigm wherein self-tests are generated through LLMs. However, Olausson et al. (2024) points out the challenge that LLMs have certain shortcomings in self-repairing their code. This paper avoids these shortcomings by proposing *functional consensus* as a reliable method of evaluation.

Program-Aided Reasoning and Agents Aside from code generation tasks, the program can be a tool that augments LLM to solve complex reasoning questions or interact with external environments. Program-of-Thought (Chen et al., 2023b) and PAL (Gao et al., 2023) prompt the model to generate a program that solves mathematical or symbolic problems. MathPrompter (Imani et al., 2023) and Chain-of-Code (Li et al., 2023a) fuse the text-based chain-of-thought with code-based program-of-thought prompting to complement each other in mathematical reasoning. Cumulative Reasoning (Zhang et al., 2024) conducts bottom-up reasoning to derive the final answer progressively. Numerous work (Sun et al., 2023; Wang et al., 2024; Yang et al., 2024) also use code as an intermediate component to bridge LLM agents with external environments.

Decompose for Complex Problems Several recent works employ decomposition to reduce the complexity of hard problems. Least-to-Most (Zhou et al., 2023) adopts a two-stage approach, which first decomposes complex problems, and then solves each sub-problem individually to tackle complex reasoning tasks. Successive Prompting (Dua et al., 2022) adopts a dynamic decomposition, iteratively breaking down problems and addressing sub-problems. Tree-of-Thought (Yao et al., 2023) breaks down complex problems into state spaces and uses tree search to solve them. Parsel (Zelikman et al., 2023) introduces decomposition to code generation tasks, taking a three-stage to break down requirements into draft and intermediate parsel programs. RepoCoder (Zhang et al., 2023) performs a retrieval in repositories to complete unfinished code one by one. Unlike these methods, FUNCODER recursively decomposes problems into a tree structure, hence gradually reduces its complexity.

5 Discussion

Limitations Our approach unleashes the potential power of functions in programming, which is advantageous on well-defined problems such as competitive programming, or program-augmented reasoning tasks. These scenarios do not however represent all use cases, such as open-ended problems or casual software development. Nevertheless, we believe that the idea of divide-and-conquer and sub-modular consensus utilized by FUNCODER can be extended to a wider range of problems, and we consider this as a future exploration.

Broader Impact While code generation is increasingly utilized in software development, Large Language Models (LLMs) are still prone to generating toxic, vulnerable, or malicious code. Such programs pose risks and should be used or executed with extra caution.

6 Conclusion

In this paper, we presented FUNCODER, a novel code generation framework that integrates the divide-and-conquer strategy with functional consensus to address complex requirements. FUNCODER had demonstrated superior performance compared to state-of-the-art methods on various benchmarks and models. Our findings highlighted the effectiveness of dynamic decomposition and functional consensus in writing complex code, which suggests that FUNCODER may have the potential to empower further improvements in code generation and other fields.

Acknowledgments

We would like to acknowledge the reviewers and chairs for their inspiring and constructive feedback. The research in this article is supported by the National Key Research and Development Project (2021YFF0901602), the National Science Foundation of China (U22B2059, 62276083). Ming Liu and Qianglong Chen are the corresponding authors.

References

- Wasi Uddin Ahmad, Md Golam Rahman Tushar, Saikat Chakraborty, and Kai-Wei Chang. AVATAR: A parallel corpus for Java-python program translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 2268–2281, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.143. URL <https://aclanthology.org/2023.findings-acl.143>.
- Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *ArXiv preprint*, abs/2108.07732, 2021. URL <https://arxiv.org/abs/2108.07732>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>.
- Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q. Feldman, Arjun Guha, Michael Greenberg, and Abhinav Jangda. Multipl-e: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Trans. Software Eng.*, 49(7):3675–3691, 2023. doi: 10.1109/TSE.2023.3267446. URL <https://doi.org/10.1109/TSE.2023.3267446>.
- Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. Codet: Code generation with generated tests. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a. URL <https://openreview.net/forum?id=ktrw68Cmu9c>.

- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *ArXiv preprint*, abs/2107.03374, 2021. URL <https://arxiv.org/abs/2107.03374>.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023b. ISSN 2835-8856. URL <https://openreview.net/forum?id=YfZ4ZPt8zd>.
- Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=KuPixIqPiq>.
- Ole-Johan Dahl, Edsger W. Dijkstra, and Charles Antony Richard Hoare. *Structured programming*, volume 8 of *A.P.I.C. Studies in data processing*. Academic Press, 1972. ISBN 978-0-12-200550-3.
- Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. Successive prompting for decomposing complex questions. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1251–1265, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.81. URL <https://aclanthology.org/2022.emnlp-main.81>.
- EbTech. How to Interpret Contest Ratings - Codeforces, 2024. URL <https://codeforces.com/blog/entry/68288>.
- Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. CodeBERT: A pre-trained model for programming and natural languages. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1536–1547, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.139. URL <https://aclanthology.org/2020.findings-emnlp.139>.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: program-aided language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICLR 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10764–10799. PMLR, 2023. URL <https://proceedings.mlr.press/v202/gao23f.html>.
- Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. Deepseek-coder: When the large language model meets programming - the rise of code intelligence. *ArXiv preprint*, abs/2401.14196, 2024. URL <https://arxiv.org/abs/2401.14196>.
- P.R. Halmos. *Naive Set Theory*. Undergraduate Texts in Mathematics. Springer New York, 1998. ISBN 9780387900926. URL <https://books.google.com.hk/books?id=x6cZBQ9qtgoC>.
- Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with APPS. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021a*. URL <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c24cd76e1ce41366a4bbe8a49b02a028-Abstract-round2.html>.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, 2021b*. URL <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html>.

- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=VtmBAGCN7o>.
- Hamel Husain, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, and Marc Brockschmidt. Codesearchnet challenge: Evaluating the state of semantic code search. *ArXiv preprint*, abs/1909.09436, 2019. URL <https://arxiv.org/abs/1909.09436>.
- Shima Imani, Liang Du, and Harsh Shrivastava. MathPrompter: Mathematical reasoning using large language models. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 37–42, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.4. URL <https://aclanthology.org/2023.acl-industry.4>.
- Xue Jiang, Yihong Dong, Lecheng Wang, Qiwei Shang, and Ge Li. Self-planning code generation with large language model. *ArXiv preprint*, abs/2303.06689, 2023. URL <https://arxiv.org/abs/2303.06689>.
- Xin Jin, Jonathan Larson, Weiwei Yang, and Zhiqiang Lin. Binary code summarization: Benchmarking chatgpt/gpt-4 and other large language models. *ArXiv preprint*, abs/2312.09601, 2023. URL <https://arxiv.org/abs/2312.09601>.
- Mohammad Abdullah Matin Khan, M. Saiful Bari, Xuan Long Do, Weishi Wang, Md. Rizwan Parvez, and Shafiq R. Joty. xcodeeval: A large scale multilingual multitask benchmark for code understanding, generation, translation and retrieval. *ArXiv preprint*, abs/2303.03004, 2023. URL <https://arxiv.org/abs/2303.03004>.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- Chengshu Li, Jacky Liang, Andy Zeng, Xinyun Chen, Karol Hausman, Dorsa Sadigh, Sergey Levine, Li Fei-Fei, Fei Xia, and Brian Ichter. Chain of code: Reasoning with a language model-augmented code emulator. *ArXiv preprint*, abs/2312.04474, 2023a. URL <https://arxiv.org/abs/2312.04474>.
- Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy V, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Moustafa-Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you! *ArXiv preprint*, abs/2305.06161, 2023b. URL <https://arxiv.org/abs/2305.06161>.
- Yujia Li, David H. Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *ArXiv preprint*, abs/2203.07814, 2022. URL <https://arxiv.org/abs/2203.07814>.
- Tianyang Liu, Canwen Xu, and Julian McAuley. Repobench: Benchmarking repository-level code auto-completion systems. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=pPjZiOUQuF>.
- Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex

- Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian J. McAuley, Han Hu, Torsten Scholak, Sébastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, and et al. Starcoder 2 and the stack v2: The next generation. *ArXiv preprint*, abs/2402.19173, 2024. URL <https://arxiv.org/abs/2402.19173>.
- Shuai Lu, Daya Guo, Shuo Ren, Junjie Huang, Alexey Svyatkovskiy, Ambrosio Blanco, Colin B. Clement, Dawn Drain, Daxin Jiang, Duyu Tang, Ge Li, Lidong Zhou, Linjun Shou, Long Zhou, Michele Tufano, Ming Gong, Ming Zhou, Nan Duan, Neel Sundaresan, Shao Kun Deng, Shengyu Fu, and Shujie Liu. Codexglue: A machine learning benchmark dataset for code understanding and generation. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c16a5320fa475530d9583c34fd356ef5-Abstract-round1.html>.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*, 2023. URL <https://openreview.net/forum?id=S37h0erQLB>.
- Meta AI. Introducing Meta Llama 3: The most capable openly available llm to date, 2024. URL <https://ai.meta.com/blog/meta-llama-3/>.
- Mistral AI. Codestral: Hello, world!, 2024. URL <https://mistral.ai/news/codestral/>.
- Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=iaYcJKpY2B_.
- Theo X. Olausson, Jeevana Priya Inala, Chenglong Wang, Jianfeng Gao, and Armando Solar-Lezama. Is self-repair a silver bullet for code generation? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=y0GJXRungR>.
- OpenAI. GPT-4 technical report. *ArXiv preprint*, abs/2303.08774, 2023. URL <https://arxiv.org/abs/2303.08774>.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.
- Nikhil Pinnaparaju, Reshith Adithyan, Duy Phung, Jonathan Tow, James Baicoianu, Ashish Datta, Maksym Zhuravinskiy, Dakota Mahan, Marco Bellagente, Carlos Riquelme, and Nathan Cooper. Stable code technical report. *ArXiv preprint*, abs/2404.01226, 2024. URL <https://arxiv.org/abs/2404.01226>.
- Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *ArXiv preprint*, abs/2307.07924, 2023. URL <https://arxiv.org/abs/2307.07924>.
- Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code. *ArXiv preprint*, abs/2308.12950, 2023. URL <https://arxiv.org/abs/2308.12950>.
- Max Schäfer, Sarah Nadi, Aryaz Eghbali, and Frank Tip. An empirical evaluation of using large language models for automated unit test generation. *IEEE Trans. Software Eng.*, 50(1):85–105, 2024. doi: 10.1109/TSE.2023.3334955. URL <https://doi.org/10.1109/TSE.2023.3334955>.
- Kensen Shi, Jacob Steinhardt, and Percy Liang. Frangel: component-based synthesis with control structures. *Proc. ACM Program. Lang.*, 3(POPL), 2019. doi: 10.1145/3290386. URL <https://doi.org/10.1145/3290386>.

- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*, 2023. URL <https://openreview.net/forum?id=vAE1hFcKW6>.
- Mohammed Latif Siddiq, Joanna C. S. Santos, Ridwanul Hasan Tanvir, Noshin Ulfat, Fahmid Al Rifat, and Vinicius Carvalho Lopes. Exploring the effectiveness of large language models in generating unit tests. *ArXiv preprint*, abs/2305.00418, 2023. URL <https://arxiv.org/abs/2305.00418>.
- Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. Adaplaner: Adaptive planning from feedback with language models. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*, 2023. URL <https://openreview.net/forum?id=rnKgbKme1t>.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288, 2023. URL <https://arxiv.org/abs/2307.09288>.
- Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better LLM agents. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024. URL <https://openreview.net/forum?id=8oJyuXfrPv>.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>.
- Yue Wang, Weishi Wang, Shafiq Joty, and Steven C.H. Hoi. CodeT5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8696–8708, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.685. URL <https://aclanthology.org/2021.emnlp-main.685>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (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*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface’s transformers: State-of-the-art natural language processing. *ArXiv preprint*, abs/1910.03771, 2019. URL <https://arxiv.org/abs/1910.03771>.
- Weixiang Yan, Yuchen Tian, Yunzhe Li, Qian Chen, and Wen Wang. CodeTransOcean: A comprehensive multilingual benchmark for code translation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5067–5089, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.337. URL <https://aclanthology.org/2023.findings-emnlp.337>.
- Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao Wang, Heng Ji, and ChengXiang Zhai. If LLM is the wizard, then code is the wand: A survey on how code empowers large language models to serve as intelligent agents. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024. URL <https://openreview.net/forum?id=8dmNOD9hbq>.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*, 2023. URL <https://openreview.net/forum?id=5Xc1ecx01h>.

- Eric Zelikman, Qian Huang, Gabriel Poesia, Noah Goodman, and Nick Haber. Parsel: Algorithmic reasoning with language models by composing decompositions. In *Thirty-seventh Conference on Neural Information Processing Systems, NeurIPS 2023*, 2023. URL <https://openreview.net/forum?id=qd9qcbVAwQ>.
- Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. RepoCoder: Repository-level code completion through iterative retrieval and generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2471–2484, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.151. URL <https://aclanthology.org/2023.emnlp-main.151>.
- Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. Cumulative reasoning with large language models. In *ICLR 2024 Workshop on Bridging the Gap Between Practice and Theory in Deep Learning*, 2024. URL <https://openreview.net/forum?id=XAAyRxTlQ>.
- Lily Zhong, Zilong Wang, and Jingbo Shang. LDB: A large language model debugger via verifying runtime execution step-by-step. *ArXiv preprint*, abs/2402.16906, 2024. URL <https://arxiv.org/abs/2402.16906>.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL <https://openreview.net/forum?id=WZH7099tgfM>.
- Ming Zhu, Karthik Suresh, and Chandan K. Reddy. Multilingual code snippets training for program translation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 11783–11790. AAAI Press, 2022. URL <https://ojs.aaai.org/index.php/AAAI/article/view/21434>.

A Appendix

In the supplementary materials, we provide the details of implementation (A.1), baseline information and settings (A.2), benchmarks (A.3), metrics (A.4), settings in the analysis (A.5), and additional experiments (A.9). We also demonstrate the example solutions of our method and baseline in Appendix B, and include all the prompts in Appendix C.

Table 5: Symbols and Glossary.

	Alias	Description
<i>(i) Symbols</i>		
$f(x)$	Function	In the programming language, a function consists of header, documentation, and its body $\{h_f, d_f, b_f\}$. A function can also be viewed as a mapping $f : D(f) \rightarrow Y$.
h_f	Function Header	Declares the function name, arguments, and return type, and is used as a signature to identify the function in a program.
d_f	Function Docstring (or Documentation)	Provides additional usage details for this function, but is optional. We encourage the model to generate docstrings to describe sub-goals precisely.
b_f	Function Body (or Implementation)	The function body contains a subroutine that describes its control flow and behavior. Functions may be invoked from within.
$f'(x)$	Partially Implemented	A provisional function structure generated by the LLM where sub-procedures are not yet implemented.
$f^*(x)$	Solved Function	A final implementation that is no longer changed and represents FUNCODER’s final comprehension and solution on the original problem.
$F = \{f_{(i)}\}$	Sampled Implementation	Functions that re-implement $f'(x)$ based on solved sub-functions, generated by models using the same input prompt.
$\text{CHILD}(f(x))$	Dependency	Functions that are used in $f(x)$. (exclude $f(x)$ itself)
T	Dependency Tree	Defined by $\text{TREE}(f, \text{CHILD}(f))$, where f is the root node of the current sub-task. Circular references are ignored.
$\mathcal{F}$	Function Composition	To implement a certain function f respecting sub-procedures as potentially reusable components.
<i>(ii) Glossary</i>		
System Test	Hidden Test	System testing is a phase where a set of previously invisible test cases are run against the submitted program to validate if the code is correct and produces the expected output for different categories of inputs.
Unit Test	Assertion	A unit test is an assertion consisting of given input and expected output, whereas in Python, it takes the form of <code>assert func(x) == y</code> .
Self-testing	-	Self-testing is an evaluation process that prompts the model to generate unit tests (assertions) to assess the correctness of the generated program.
AlphaCode-like Clustering	-	AlphaCode proposed a clustering process that elects candidate program from a number of samples, recognizing programs that produce exactly identical outputs as equivalent, and picks one program from the largest cluster.

A.1 Implementation Details

Models We access the OpenAI models GPT-3.5 (gpt-3.5-turbo-0613), GPT-4 (gpt-4-1106-preview) and GPT-4o mini (gpt-4o-mini-2024-07-18) through Azure OpenAI. Weights of community models Llama3 (Meta-Llama-3-8B-Instruct), Codestral (Codestral-22B-v0.1), StableCode (stable-code-instruct-3b), CodeLlama (CodeLlama-34b-Instruct-hf) and StarCoder2 (starcoder2-15b-instruct-v0.1)

are downloaded from HuggingFace (Wolf et al., 2019) and served over an OpenAI-like API on a single A100-80G GPU under BF16 precision with vLLM (Kwon et al., 2023).

Divide We instruct the model to write the current function and introduce new functions with clearly defined sub-goals. The prompt C.2 for the *divide* process includes two examples: one example needs to involve new functions that are left unimplemented; and another where the sub-goal is simple enough that no further decomposition is necessary. The model generates a Python code block with a temperature of 0.2, and the code block will be extracted to represent a tree of functions with new functions as the children of the current. We require that any new sub-function do not refer to existing functions, to avoid circular references. This generation process will be attempted at most 3 times until any valid code with a proper signature is found in the output. FUNCODER then traverses the function tree via depth-first search and restricts the max depth of the tree to 6.

Conquer We apply the composition of sub-functions to rewrite the parent function after all sub-functions have been fully decomposed. Code for sub-functions is made visible to the LLM, which is requested to rewrite the current function with a 1-shot demonstration (C.3). With functional consensus applied, the model samples multiple implementations with a temperature of 0.8, and the one that reaches consensus will be kept for further bottom-up processing.

Functional Consensus The functional consensus is applied in the conquer stage. Formally, Consensus@k samples k-1 implementations in the conquer stage, and reuses the one produced in the divide stage, resulting in a set F of k candidate programs. Then we prompt the model with 1-shot (C.4) to generate potential inputs X for the given function and use them to feed and execute the program. As described in Eq 2, when two functions output the same value in a given input, they will both add 1 point to the overall similarity. A thrown exception or timeout during execution assigns -100 points to the candidate as it indicates potentially problematic code. Similar to self-testing methods, we also leverage the example input/output at the root node to filter out candidates that have wrong functionality. Finally, the one candidate with maximum scores over all inputs is selected, as it reaches consensus with other implementations.

Hierarchical Code Interpreter Divide-and-conquer represents the problem hierarchy through structured code. To gain insights of this information, we design an interpreter that syntactically parses the generated output and organizes them into a graph of functions. We are thus able to decompose complex tasks by representing sub-goals through the connections of multiple functions. LLMs may produce vulnerable code even if prompted by trusted inputs, making direct execution or `eval()` on generated code especially hazardous. Our framework addresses this with the use of a sandboxed environment to contain untrusted code execution, preventing the LLM from hanging up or even breaking the system.

A.2 Baseline Details

Standard Prompting conducts one-time generation and directly output the entire code or final results. In code generation tasks, we use a 1-shot prompting setting with 0.3 temperature. For MATH, we sample 1 question-answer pair per subject in the train set, resulting in a 7-shot prompt, and run self-consistency (Wang et al., 2023) with consistency@5 and temperature 0.7.

CodeT (Chen et al., 2023a) samples multiple code solutions X and unit-tests Y . A unit test is an assertion consisting of given input and expected output, whereas in Python it takes the form of "`assert func(x) == y`", CodeT then checks the programs over self-tests and divides the functions into sets; the score of such a set is defined as the number of functions within multiplied by the number of succeeded tests. Finally, CodeT selects the function with the most agreement (in the biggest set). Similar to the setting of FUNCODER, we sample 11 candidate solutions with 0.8 temperature.

AlphaCode-like Clustering is introduced with the model AlphaCode (Li et al., 2022), and samples multiple code solutions X . A fine-tuned model is used to generate test inputs upon which programs are evaluated. Programs are then clustered by whether the outputs are (exactly) identical, in which the largest group was selected. The final result comes from this largest group, where any result within would suffice. In our ablation study, we use similar settings to FUNCODER, sampling 11 candidate solutions with 0.8 temperature and generate sample inputs likewise.

Parsel (Zelikman et al., 2023) consists of three generation stages: high-level sketch, Parsel program, and final program. The Parsel program is an intermediate representation of code that describes and organizes program structure. We report the result of HumanEval with GPT-4 from the original paper.

Reflexion (Shinn et al., 2023) is a closed-loop agent system that generates unit tests and iteratively refines the program based on the self-test feedback. The results for GPT-4 on HumanEval and MBPP are reported in the original paper. Based on officially released code³, we test results with GPT-3.5 and community models under the reflexion strategy with `max_iters=2` and `Pass@1`. For the xCodeEval benchmark, as it is judged through standard input/output, we wrap the standard input into function arguments and obtain the return value as the output in the form of `"def main(input_str: str) -> str"`, and the sample input/output are also transformed to visible tests for reflexion process.

MetaGPT (Hong et al., 2024) employs a multi-agent strategy that assigns roles and encodes human-like software development procedures. The scripts for reproducing the results were not made public as of this paper was completed. Therefore, we include the original result for GPT-4 on the HumanEval dataset under the with feedback setting.

LDB (Zhong et al., 2024) segments programs into basic blocks and tracks the values of intermediate variables after each block throughout runtime execution, allowing large language models to verify the correctness of smaller code units. We adopt the results as-is reported in the paper.

Chain-of-Thought Prompting (Wei et al., 2022) generates step-by-step reasoning leading to the final output answer. The solution is formatted in $\LaTeX$, and use `\boxed` to mark the final answer. We sample 1 shot per subject in the MATH train set, resulting in a 7-shot demonstration, and running with consistency@5 and a temperature of 0.7.

Program-of-Thought (Chen et al., 2023b) utilizes the coding ability in LLMs to generate programs rather than text-based solutions for reasoning tasks. In MATH, we hint the model with 1-shot prompting to generate a `solution()` function that returns the final answer to the problem. The program is then executed in a Python environment and obtains the return value. If an exception is thrown during execution, the model will try to regenerate a new program until it succeeds or reaches 3 attempts. Similar to CoT, Program-of-Thought samples 5 programs at a temperature of 0.7 and votes the final result.

Self-Refine (Madaan et al., 2023) iteratively prompts the model to give feedback and refine the generated code based on it. Self-refine does not incorporate self-tests, and the refinement is conducted solely on model feedback. In our preliminary study on HumanEval, this feedback is weak and cannot improve performance. However, in MATH, the solution program can be executed without the need for generated assertions. Thus, we extend the self-refine to capture the runtime error trace as feedback and refine the code until it can run or exceed 3 retries.

Cumulative Reasoning (Zhang et al., 2024) starts from decomposing the input problem into propositions and conducts bottom-up reasoning until the final answer can be concluded. The results for Cumulative Reasoning are reported in the original paper under with code setting.

A.3 Benchmark Details

Table 6: Overview and details of HumanEval, MBPP, xCodeEval, and MATH dataset.

	HumanEval	MBPP	xCodeEval	MATH
Task Attribute	Code Generation	Code Generation	Programming Contest tags, difficulty	Mathematical Reasoning subject, level
Metric	Pass@1	Pass@1	Pass@1	EM-gpt
# Sample (original)	164	427	7,635	5,000
# Sample (ours)	164	200	500	500
Entry func	variant	variant	main()	solution()
Input	arguments	arguments	standard input	n/a
Output	return	return	standard output	return
# Examples Tests	~2.8	0	~2.1	n/a
# System Tests	~8.1	~3.1	51.1	n/a

³GitHub: noahshinn/reflexion

HumanEval (Chen et al., 2021) is a hand-crafted programming dataset designed to evaluate a model’s code generation capability. It consists of 164 instances involving programming skills in language comprehension, reasoning, algorithms, and simple mathematics. The problem contains 2.8 sample inputs and outputs on average in the function document, which can be leveraged to provide additional guidance for the LLM to select or self-improve the programs. We conduct experiments on all 164 instances using accuracy (Pass@1) as the evaluation metric. The details of the Pass@1 metric are described in Appendix A.4.

MBPP (Austin et al., 2021) consists of fundamental Python programming problems, with a total of 974 examples covering Python programming basics, standard library usage, and related assessment. Following Shinn et al. (2023), we adopt the mbpp-typed split from MultiPL-E (Cassano et al., 2023) and sample 200 instances, using Pass@1 as the metric. The original prompt⁴ from MBPP includes all hidden tests in the input problem, which may cause label leakage when using these tests to refine or select programs. To ensure a fair comparison, MultiPL-E removes the test information in the prompt.

xCodeEval (Khan et al., 2023) is a competition-level multilingual and multitask benchmark consisting of 17 programming languages. xCodeEval collects 25 million openly available samples from codeforces.com, a platform for competitive programming. The data we use include problem descriptions in `problem_descriptions.jsonl` and system tests from `unittest_db.json` which consists of 7,635 competition problems and averaged 51.1 tests per problem. Note that the tests in xCodeEval are crawled, some of them are incomplete due to the website context limit (they end with an ellipsis and the further content is missing); we filter out problems having invalid test cases. Based on the CodeForces Rating (EbTech, 2024), we categorize the problems by their difficulty: Easy (≤ 1200), Mid (1200-1599), Hard (1600-1999), and Expert (≥ 2000). We sample 500 problems from the full split with the basic filter rule mentioned above, resulting in Table 7. The CodeForces problem has a different input/output style compared to HumanEval and MBPP; it scans input from Standard Input and prints the answer to Standard Output. Therefore, we judge the program on system tests using a CodeForces-style judger and use Pass@1 (the program must pass all system tests) as the evaluation metric.

MATH (Hendrycks et al., 2021b) is a challenging competition-level mathematical reasoning dataset, with problems and solutions formatted in $\text{\LaTeX}$. It covers seven categories: Prealgebra, Algebra, Number Theory, Counting & Probability, Geometry, Intermediate Algebra, and Precalculus. The original test set of MATH consists of 5000 samples, and we randomly sampled 500 problems as shown in Table 7. In addition to text-based reasoning, writing programs is another promising way to solve mathematical problems. These methods involve writing a `main()` or `solution()` function, and executing the program to obtain the final answer. Through experiments on MATH, we aim to demonstrate that FUNCODER can enhance LLM’s ability to address complex mathematical problems through programming.

Table 7: Number of test samples in (a) xCodeEval difficulty, (b) MATH level, (c) MATH subject.

Difficulty	Ours	Original	Level	Ours	Original	Subject	Ours	Original
Easy	178	1428	Level 1	39	437	Prealgebra	82	871
Mid	112	1319	Level 2	90	894	Algebra	139	1187
Hard	87	1453	Level 3	108	1131	Number Theory	45	540
Expert	118	3289	Level 4	116	1214	Counting & Probability	47	474
n/a	5	146	Level 5	147	1324	Geometry	42	479
Total	500	7635	Total	500	5000	Intermediate Algebra	90	903
						Precalculus	55	546
						Total	500	5000

A.4 Metrics

Pass@k When a program is passed (or accepted), it means that the program must pass all system tests without errors and within the time limit. In our experiments, we set the time limit to 2.5 seconds. Pass@k judges k independent programs, and if any of them can pass, the result will be 1.

⁴Original MBPP prompt: <https://github.com/google-research/google-research/tree/master/mbpp>

In most of our experiments, we use Pass@1 as the metric, as it reflects the accuracy of the method framework achieved without feedback from humans. Pass@k, on the other hand, is equivalent to filtering programs through hidden, human-annotated test labels.

EM-GPT The ground truth label in MATH is written in $\text{\LaTeX}$, and the accuracy between labels and model predictions cannot be directly calculated through exact-match (EM). MATH provides a judge program⁵ that preprocesses LaTeX syntax and check whether two disambiguated strings are equal. However, this is insufficient for evaluating LaTeX-formatted labels with variant program outputs. We follow the evaluation criteria from previous work (Zhang et al., 2024), using GPT-4 to assess the consistency between predictions and ground truths, with prompt shown in C.6.

A.5 Details of Analysis

Details of Preliminary Analysis on Self-testing (Figure 3.a) The preliminary study is conducted on the HumanEval dataset, which includes system tests S to evaluate the accuracy of the program, as well as one human-annotated canonical solution c . For each question, we: (1) Obtain one solution program p from Standard Prompting. (2) Prompt the model to generate 7 self-tests T based on the question and entry function. The self-test is in the form of the unit test `assert f(x) == y`. We then judge the generated program p and canonical solution c over the self-tests T and system tests S . Formally, a pair (x, Y) is used to identify whether program x passes test Y . Where (p, S) indicates that the program can pass the system tests, demonstrating its correctness. And $\neg(c, T)$ means the canonical solution can not pass self-tests, suggesting that the tests generated by model could be wrong. The self-test results on generated programs are first evaluated and divided into two classes: self-test passed or failed. If the self-test passes, the self-improvement methods will stop iteration and pick this program as a final result. The next step is to determine whether the program can pass system tests. If the self-test fails, it indicates that there could be an error in the program or test itself. In this case, the correctness of the program is checked using final tests (p, S) and the correctness of the unit test by canonical program (c, T) . The results on GPT-3.5 and StableCode are shown in Figure 3 and detailed explanations about these conditions can be found in Table 8.

Table 8: Explanation on how we classify cases in self-testing preliminary study.

Class	Subclass	Condition	Explanation
self-test passed	final passed	$(p, T) \wedge (p, S)$	The self-test result is consistent with the final judge. However, self-testing methods cannot improve performance in this case, as the program from the baseline (Standard Prompt) is already correct.
	final failed	$(p, T) \wedge \neg(p, S)$	Self-test is too weak to detect errors in the program, there could be edge cases that not been considered.
self-test failed	program wrong	$\neg(p, T) \wedge \neg(p, S) \wedge (c, T)$	This is a good example that self-testing detects errors in the program. Feedback from the test will be used to select or refine the solution.
	unit-test wrong	$\neg(p, T) \wedge (p, S) \wedge \neg(c, T)$	Bad case, the self-test produced an error result and filtered out a correct solution. Continuously revising the code for this test will lead to a performance downgrade.
	both wrong	$\neg(p, T) \wedge \neg(p, S) \wedge \neg(c, T)$	The model is unable to generate a correct solution or test cases. Refining the program over faulty test samples will not lead to the correct answer.
	-	$\neg(p, T) \wedge (p, S) \wedge (c, T)$	In the event of self-test failure, there must have been at least one error in either program or tests, so this condition should never occur .

Details of Ranking Strategy Comparison (Figure 3.b) We obtain 11 candidate programs from FUNCODER on HumanEval with GPT-3.5 and rank them through three strategies. This ensures that the same candidate set is used for a fair comparison. An effective ranking strategy should prioritize

⁵math_equivalence: https://github.com/hendrycks/math/blob/main/modeling/math_equivalence.py

placing correct programs at the forefront and filter out those with errors. Thus, we measure the effectiveness by computing Pass@k results on the top-k-ranked programs selected by each strategy. The Pass@11 result serves as an upper bound as it uses all programs to compute the pass rate.

How We Count Frequently Used Functions in MATH (Table 4) In the mathematical reasoning experiments, we used a subset of 500 items from the MATH test set, with an average of 71.4 questions per subject. However, it is not very confident to represent common functions from only 71.4 programs. Therefore, we sample 3000 problems from the MATH test set for this experiment and run the *divide-only* setting of FUNCODER on them. Then, the occurrence of sub-functions is counted based on their names after extracting the function nodes of code trees for each category.

A.6 Detailed Explanation of Algorithm

We hereby provide a detailed explanation of FUNCODER algorithm works, with respect to Algorithm 1 from Figure 2 (a copy is included below for simple reading). As mentioned, FUNCODER is a recursive process following a DFS pattern. We use square brackets (e.g. [L1]) below to denote line numbers in the pseudocode.

Algorithm 1 FUNCODER procedure

Require: Entry func, $f_{\text{root}} = \{h_{\text{root}}, d_{\text{root}}, \phi\}$
Require: Large language model, LLM

```

1: function FUNCODER( $f_{\text{cur}}$ )
2:   — Divide —
3:    $f'_{\text{cur}}, \{f_i\} \leftarrow \text{EXTRACT}(\text{LLM}(f_{\text{cur}}))$ 
4:   for  $f_i \in \{f_i\}$  do
5:     if  $b_i$  is NOT IMPLEMENTED then
6:        $f_i^* \leftarrow \text{FUNCODER}(f_i)$  ▷ recursion
7:     end if
8:      $\text{ADDCHILD}(f_{\text{cur}}, f_i^*)$ 
9:   end for
10:  — Conquer —
11:   $F_{\text{cur}} \leftarrow \text{SAMPLE}(\text{LLM}(f'_{\text{cur}}, \text{CHILD}(f_{\text{cur}}))$ 
12:   $f_{\text{cur}}^* \leftarrow \text{FUNCONSENSUS}(F_{\text{cur}})$ 
13:  return  $f_{\text{cur}}^*$ 
14: end function
15: return FUNCODER( $f_{\text{root}}$ ) ▷ starts from root

```

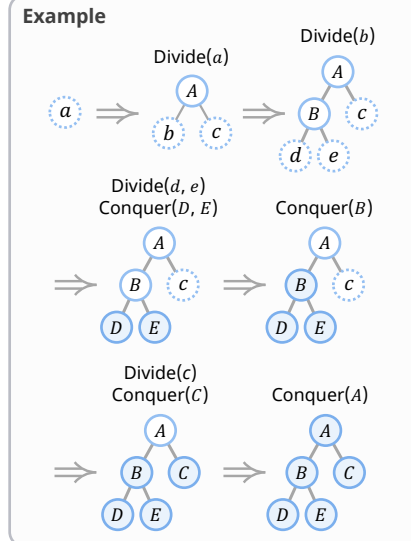


Figure 5: Left: Algorithm for FUNCODER. Right: Decomposition example of $A[B[DE]C]$.

FUNCODER [L1], when solving each function f , first performs the *Divide* stage [L3-L9], where the LLM initially writes the function and identifies some potential sub-problems, represented as sub-function stubs (e.g., `def f(xs: list[int]) -> int`) [L3]. In this process, we identify the sub-problems of the current problem, thereby understanding the dependency between functions.

For each decomposed sub-problem $g_1, g_2, \dots$, we recursively use FUNCODER to obtain the final implementation G_i for that sub-problem [L5-L8]. This G_i shall replace the previously incomplete subfunction stub signature in the final program.

Now that all sub-problems of f are implemented, we move on to the *Conquer* stage [L11-13] to complete the larger problem. By combining the signature f and the final implementations $G_1, G_2, \dots$ of sub-problems, we generate the complete implementation F [L11] and return it [L13].

We hierarchically describe how this algorithm works in detail by combining it with the example given in the right half of Figure 5.

```

[a.1] FunCoder(a)
| [a.3] LLM(a) -> A, {b, c}      # divide
|--[b.1] FunCoder(b)
| | [b.3] LLM(b) -> B, {d, e}    # divide

```

```

| |--[d.1] FunCoder(d)
| | | [d.3] LLM(d) -> D, {}      # divide
| | +--[d.13] return D          # nothing to conquer
| |--[e.1] FunCoder(e)
| | | [e.3] LLM(e) -> E, {}      # divide
| | +--[e.13] return E          # nothing to conquer
| | [b.11] LLM(B, {D, E}) -> B*  # conquer
| +--[b.13] return B*
|--[c.1] FunCoder(c)
| | [c.3] LLM(c) -> C, {}        # divide
| +--[c.13] return C            # nothing to conquer
| [a.11] LLM(A, {B, C}) -> A*    # conquer
+--[a.13] return A*              # final result

```

A.7 Token-cost Complexity

Example We use the example from Figure 5, where the final program consists of 5 functions $A[B[D,E],C]$, and A serves as the entry to the program. Here we respect the aforementioned notations, and further use the lower-case letter a to represent the number of stub tokens, upper-case A to represent the number of result tokens, for the function A , and other functions likewise. Let $N = A + B + C + D + E$ be the token-cost complexity of the final result.

In *Standard*, the code is only generated once to complete the given stub. We use parentheses to represent the order of LLM calls in a full process.

```

(1)          a -> A B C D E
input tokens = a
output tokens = A + B + C + D + E
overall      = O(N)

```

In each step of the *FUNCODER/Divide* stage, the to-be-implemented function will serve as the context. The function (stub) will be implemented and sub-function stubs are to be declared.

```

(1)          a -> A b c
(2)          b -> B d e
(3)          d -> D
(5)          e -> E
(8)          c -> C
input tokens = a + b + c + d + e
output tokens = A + b + B + c + C + d + D + e + E < 2N
overall      = O(N)

```

In every *FUNCODER/Conquer*, the context shall include the current function's definition and finalized implementations of sub-functions. The output is the re-implemented current function.

```

(4)          d -> kD
(6)          e -> kE
(7)          b D E -> kB
(9)          c -> kC
(10)         a B C -> kA
input tokens = a + b + B + c + C + d + D + e + E < 2N
output tokens = kA + kB + kC + kD + kE = kN
overall      = O(kN)

```

These stages in all bring *FUNCODER*'s total token consumption to strictly $O(kN)$ for every problem.

Token Complexity of *FUNCODER* is $O(kN)$ Define N as the token length of the final program, which is correlated to the inherent complexity of the problem, and define k as the number of sampled candidates. We first explain in detail that the worst-case token cost of *FUNCODER* is $O(kN)$:

- The naive *Standard* method should naturally generate $O(N)$ tokens. Sampling-based baselines like *CodeT* cost $O(kN)$ tokens.
- FunCoder goes through the *Divide* stage and the *Conquer* stage for each of the functions.
- Based on the current function, *Divide* generates an implementation of itself and stubs for sub-functions. Within this stage, each function would appear at most once in input and twice in output. All *Divide* stages consume no more than $3N$ tokens.
- *Conquer* regenerates the parent function based on its stub and all finalized sub-functions. Herein each function will appear at most twice in input, and sampled k times in output. If $k = 1$, consensus is implicitly disabled. All *Conquer* stages shall consume at most $(k + 2)N$ tokens.

So FunCoder requires no more than $(k + 5)N$ tokens in input-output, making its token consumption $O(kN)$ even at worst-case, aligning with other sampling-based methods such as CodeT and AlphaCode-like clustering. Furthermore, when sampling is disabled ($k = 1$), our method has a token consumption of $O(N)$, which also aligns with the vanilla *Standard* method.

A.8 Discussion About Functional Consensus

This section focuses on why *functional consensus* might enhance the correctness of programs and how it differs from other consistency-based methods. Self-consistency (Wang et al., 2023) is widely employed in the realm of LLM reasoning. It samples multiple sets of answers and uses voting to select the most consistent result, where the answers typically consist of named entities, choice options, or numbers. However, this approach faces challenges when voting on sampled programs, as programs describe executable logic instead of data, making it unobvious to determine whether two programs are equivalent just from the looks.

When it comes to picking programs, *functional consensus* in FUNCODER looked beyond the literal symbols and used a different approach. It uses inputs and execution results to compare behavioral differences among programs. There have been similar methods, such as the strict clustering approach in AlphaCode (Li et al., 2022), which samples a set of program inputs and then clustering programs with identical outputs into the same group. The final program is then selected from the largest cluster.

However, the idea of grouping programs by the ‘identicalness’ of outputs is not without fallacies, since programs rarely specialize in solving one single irreducible problem – they deal with a variety of inputs, conditions and mysterious cases. The result of this, where different solutions could have many common behaviors and some distinct behaviors, is referred to as the term ‘special-case similarity’ in the FrAngel paper (Shi et al., 2019). We consider a correct program solution which has multiple ‘special-case similar’ programs that are partially correct in different ways, for example:

- One program behaving correctly on the general case (almost all) but missed a few edge cases
- Another program got one edge case correct but didn’t manage to deal with the general case
- Yet another program got all edge cases correct but crashed on the general case
- A buggy program that behaves correctly on all available test cases but none of these tests trigger the bug (literally test coverage problem)
- And many programs that turn out to be frenzy mixtures of all the above

If we had a pool of programs that contained the fully-correct program and an assortment of other programs that respected certain cases of the problem as aforementioned, it’d be obvious that the fully-correct would be decently ‘special-case similar’ to the rest of the programs, for their similar behavior on inputs. These execution outputs are programmatically obtained and automatically compared against each other without any human intervention or LLM calling required, the process of which sits at the core of our *functional consensus* algorithm.

Therefore, with *functional consensus*, where the solutions with common behaviors are promoted, we could intuitively expect the result to be a higher likelihood of a fully-correct program. Provided below is a hypothetical example demonstrating why *functional consensus* prevails:

Example Consider the problem of finding all square roots in the complex domain of a non-zero real number (stored in `float32`). To get the answer right for all inputs, the function must consider 2

cases: A) non-zero numbers have 2 square roots; B) square roots of negative numbers are imaginary. 10 candidate functions are sampled as below:

- 5 results (a_1, a_2, a_3, a_4, a_5) only considered case A and got just positive inputs right. For negative numbers, they literally gave up and crashed.
- 3 results (b_6, b_7, b_8) remembered to consider case B, gave 2 imaginary results for negative numbers, but forgot to do the same for positive numbers, returning only 1 square root therein.
- Only 2 results (c_9, c_{10}) considered all cases and returned correct results for all inputs.

If we pick the program through ‘clustering’, the final result would be one of the 5 results ($a_1, a_2, \dots, a_5$) that only considered case A, which is evidently not the correct solution. But with *functional consensus*, the final result is vastly different, since we consider the similarity between the functions based on their behavior on different inputs. Without loss of generality, suppose that there are 2 test inputs 4.0, −9.0, one for each of the 2 cases. We calculate the similarity as follows:

- Programs a_i got only $\sqrt{4.0} = [2.0, -2.0]$ right so each program here are similar with programs ($a_1, a_2, a_3, a_4, a_5, c_9, c_{10}$), scoring 7 points.
- Since b_j only went well with $\sqrt{-9.0} = [3.0i, -3.0i]$, programs here only score 5 points for case B with the ones ($b_6, b_7, b_8, c_9, c_{10}$).
- Each program in c_k gets 7 points for $\sqrt{4.0}$ with ($a_1, a_2, a_3, a_4, a_5, c_9, c_{10}$), and gets 5 points for $\sqrt{-9.0}$ with ($b_6, b_7, b_8, c_9, c_{10}$). Totals to 12 points.

The final result apparently leaned towards c_k as the correct solution, even if their outputs as a whole weren’t even half as much as a_i is. Through this example, we illustrate that *functional consensus* has the potential to identify the correct samples even at their minority, outperforming other methods such as self-consistency or clustering.

A.9 Supplementary Results

Token Usage We provide token usage results in Table 9 for FUNCODER and baseline methods on the HumanEval dataset with the GPT-3.5 model, whereas usage results on other datasets are provided in Table 10. We report the average token usage per problem. The token usage is computed through the sum of prompt tokens and completion tokens returned by OpenAI API chat completion call⁶. For LDB, we report their token usage in the original paper (Zhong et al., 2024).

Table 9: Token usage for different settings of FUNCODER and baseline methods on HumanEval, all evaluated on GPT-3.5-turbo. The LDB results are reported in the original paper. The main setting for LDB and FUNCODER is bolded.

Method	Setting	Pass@1	Tokens			
			Min.	Max.	Avg.	Med.
Standard	One-time	68.3	648	1477	886.7	861
CodeT	One-time + Self-Test@11	81.1 (+12.8)	2298	9645	4479.1	4166
Reflexion	maxiter=2	69.5 (+1.2)	416	4906	1416.1	754
LDB (<i>reported</i>)	line-level	80.5 (+12.2)	-	-	24K	-
	block-level	82.9 (+14.6)	-	-	23K	-
	function-level	79.9 (+11.6)	-	-	27K	-
FUNCODER	One-pass	72.6 (+4.3)	826	3489	1233.7	1132
	Two-pass	78.7 (+10.4)	2197	8406	3343.2	3078
	Two-pass + Consensus@5	83.5 (+15.2)	2455	9432	4040.9	3800
	Two-pass + Consensus@11	85.4 (+17.1)	3015	13850	5402.0	5166
FUNCODER (<i>ablation</i>)	Two-pass + Self-Test@11	80.5 (+12.2)	2967	13758	5408.3	5184
	Two-pass + Clustering@11	75.0 (+6.7)	3044	9958	5070.7	4888

⁶<https://platform.openai.com/docs/guides/text-generation/managing-tokens>

Table 10: Token usage of FUNCODER and baseline methods on other datasets, i.e. MBPP, xCodeEval and MATH. Results are evaluated on GPT-3.5-turbo.

Dataset	Method	Pass@1	Tokens			
			Min.	Max.	Avg.	Med.
MBPP	Standard	72.0	577	2102	744.9	717.0
	CodeT	76.0 (+4.0)	2232	8172	2945.3	2866.0
	Reflexion	72.5 (+0.5)	391	3379	1205	569.5
	FUNCODER	78.5 (+6.5)	3462	13229	5049.9	4644.0
xCodeEval	Standard	20.2	1051	3343	1599.5	1530.0
	CodeT	23.2 (+3.0)	2264	9245	3937.4	3733.0
	Reflexion	20.6 (+0.4)	2977	1003222	401767.3	328591.5
	FUNCODER	31.4 (+11.2)	4883	53225	10559.7	8927.0
MATH	PoT	41.0	551	5867	953.0	835.0
	FUNCODER	54.0 (+13.0)	2622	30139	7075.5	5666.5

Full Results for Code Generation We provide results for all conducted experiments on code generation benchmarks in Table 11. Our method consistently improves the baseline on community models by averaging 11% on MBPP and 150% on xCodeEval. It is worth noting that small models have a tendency to have low pass rates on competition problems, leading to a relatively higher randomness, therefore we run 3 experiments and report the median result.

Full Results for MATH The MATH dataset divides the problems into five levels of difficulty. The difficulty distribution of our test set can be found in Table 7. We report the average accuracy of FUNCODER and other methods for each math subject in Table 12 and results for each level in Table 13. The results of Cumulative Reasoning are obtained from the original paper (Zhang et al., 2024). Experiment results demonstrate that our method consistently enhances the model’s reasoning ability across all levels of MATH.

Table 11: Results for Code Generation. We report Pass@1 as evaluate metric. Results from the original paper are underlined, and the best results are bold.

Model	Method	HumanEval		MBPP		xCodeEval				
		Pass@1	$\Delta \uparrow$	Pass@1	$\Delta \uparrow$	Easy	Mid	Hard	Expert	All
GPT-3.5	Standard	68.3	-	72.0	-	44.4	15.2	4.6	0.0	20.2
	CodeT	81.1	+12.8	76.0	+4.0	50.6	16.1	8.0	0.0	23.2
	Reflexion	69.5	+1.2	72.5	+0.5	44.4	17.0	5.7	0.0	20.6
	LDB	<u>82.9</u>	+14.6	<u>76.0</u>	+4.0	-	-	-	-	-
	FUNCODER	85.4	+17.1	78.5	+6.5	62.4	29.5	11.6	0.0	31.4
GPT-4	Standard	82.9	-	73.5	-	68.5	39.3	19.5	1.7	37.4
	Parsel	<u>85.0</u>	+2.1	-	-	-	-	-	-	-
	CodeT	90.9	+8.0	77.0	+3.5	76.4	51.8	21.8	3.4	44.0
	Reflexion	91.0	+8.1	<u>77.1</u>	+3.6	71.3	41.1	19.5	2.5	38.6
	MetaGPT	<u>85.9</u>	+3.0	-	-	-	-	-	-	-
	FUNCODER	94.5	+11.6	79.5	+6.0	83.1	58.0	26.4	3.4	48.6
GPT-4o mini	Standard	87.3	-	76.0	-	65.7	44.6	9.2	0.0	35.4
	CodeT	90.9	+3.6	75.5	-0.5	71.9	49.1	16.1	0.0	39.6
	FUNCODER	91.5	+4.2	77.5	+1.5	72.5	52.3	11.5	0.0	39.8
Llama3 _{8b}	Standard	61.6	-	60.5	-	9.0	1.8	0.0	0.0	3.6
	CodeT	68.9	+7.3	61.5	+1.0	12.4	0.0	0.0	0.0	4.4
	FUNCODER	79.7	+18.1	62.5	+2.0	22.0	0.9	0.0	0.0	8.0
Codestral _{22b}	Standard	79.3	-	68.5	-	27.5	4.5	2.3	0.0	11.4
	CodeT	86.0	+7.3	74.0	+5.5	34.8	7.1	3.4	0.0	14.8
	FUNCODER	89.0	+9.7	74.5	+6.0	49.4	15.2	3.4	0.0	22.0
StableCode _{3b}	Standard	61.0	-	51.5	-	7.3	0.9	0.0	0.0	2.8
	CodeT	75.0	+14.0	57.5	+6.0	11.2	1.8	0.0	0.0	4.6
	FUNCODER	81.0	+20.0	63.5	+12.0	13.5	4.5	1.1	0.0	6.2
CodeLlama _{34b}	Standard	43.9	-	53.5	-	2.3	0.0	0.0	0.0	0.8
	CodeT	55.5	+11.6	56.5	+3.0	10.1	0.0	0.0	0.0	3.6
	FUNCODER	66.5	+22.6	58.5	+5.0	10.2	0.0	0.0	0.0	3.6
StarCoder2 _{15b}	Standard	59.8	-	64.5	-	18.0	0.9	2.3	0.0	7.2
	CodeT	70.7	+10.9	66.0	+1.5	13.5	0.9	0.0	0.0	5.0
	FUNCODER	78.7	+18.9	70.0	+5.5	29.2	4.5	0.0	0.0	11.6

Table 12: Experimental results on MATH, a competition-level mathematical reasoning benchmark. Best results are in bold. Text-based reasoning methods are denoted with $\dagger$, while others use program-aided reasoning. We report both overall results and results in seven subjects: *Prealgebra*, *Algebra*, *Number Theory*, *Counting & Probability*, *Geometry*, *Intermediate Algebra*, and *Precalculus*.

Model	Method	<i>Prealg.</i>	<i>Alg.</i>	<i>NT</i>	<i>Prob.</i>	<i>Geo.</i>	<i>InterAlg.</i>	<i>Precalc.</i>	Overall
GPT-3.5	Standard $\dagger$	62.2	37.4	20.0	29.8	31.0	24.4	21.8	34.6
	CoT $\dagger$	59.8	51.1	28.9	29.8	28.6	26.7	30.9	40.0
	PoT	68.3	50.4	33.3	48.9	21.4	18.2	29.1	41.0
	Self-Refine	74.4	49.6	48.9	57.4	28.6	35.6	36.4	48.6
	FUNCODER	76.8	61.2	55.6	59.6	34.1	36.0	41.8	54.0
GPT-4	Standard $\dagger$	81.7	82.7	71.1	72.3	59.5	46.7	47.3	68.2
	CoT $\dagger$	84.1	87.1	62.2	68.1	45.2	48.9	54.5	68.6
	PoT	79.3	80.6	75.6	72.3	50.0	47.8	58.2	68.2
	Self-Refine	82.9	82.0	77.8	76.6	54.8	55.6	63.6	72.2
	CR	86.6	86.3	88.7	71.1	53.7	51.5	51.8	72.2
	FUNCODER	89.0	92.8	82.2	83.0	59.5	63.3	56.4	78.2
GPT-4o mini	Standard $\dagger$	79.3	83.5	75.6	87.2	47.6	57.8	56.4	71.8
	CoT $\dagger$	90.2	95.7	82.2	68.1	50.0	61.1	61.8	77.2
	PoT	80.5	84.2	77.8	72.3	50.0	60.0	50.9	71.0
	Self-Refine	79.3	83.5	75.6	87.2	47.6	57.8	56.4	71.8
	FUNCODER	81.7	83.5	80.0	80.9	59.5	60.0	54.5	73.2
Llama3 _{8b}	CoT $\dagger$	56.1	47.5	31.1	34.0	40.5	14.4	38.2	38.6
	PoT	67.1	32.4	24.4	34.0	16.7	21.1	18.2	32.6
	FUNCODER	67.9	45.7	51.1	53.2	19.0	37.8	30.9	45.0
Codestral _{22b}	PoT	70.7	56.1	46.7	44.7	21.4	26.7	30.9	45.6
	FUNCODER	81.7	61.9	46.7	55.3	28.6	45.6	38.2	54.8
StableCode _{3b}	PoT	20.7	14.4	17.8	25.5	4.8	8.9	9.1	14.4
	FUNCODER	46.3	30.2	20.0	29.8	4.8	20.0	18.2	26.6
CodeLlama _{34b}	PoT	35.5	26.1	15.0	16.7	0.0	5.5	33.3	15.2
	FUNCODER	44.8	46.1	37.8	34.1	13.6	24.6	37.5	24.4
StarCoder2 _{15b}	PoT	46.3	29.5	28.9	25.5	21.4	27.8	23.6	30.2
	FUNCODER	72.0	39.6	40.9	46.8	23.8	28.1	27.3	40.8

Table 13: Full results of each method at different levels of MATH. The best results are in bold. Text-based reasoning methods are denoted with [†], while others use program-aided reasoning.

Model	Method	Level 1	Level 2	Level 3	Level 4	Level 5	Overall
GPT-3.5	Standard [†]	61.5	51.1	43.5	25.9	17.7	34.6
	CoT [†]	76.9	48.9	50.9	33.6	21.8	40.0
	PoT	61.5	51.1	56.5	33.6	24.1	41.0
	Self-Refine	84.6	61.1	65.7	32.8	31.3	48.6
	FUNCODER	84.6	65.9	68.5	43.1	37.4	54.0
GPT-4	Standard [†]	89.7	85.6	83.3	55.2	51.0	68.2
	CoT [†]	94.9	81.1	77.8	64.7	50.3	68.6
	PoT	94.9	80.0	74.1	63.8	53.1	68.2
	Self-Refine	94.9	81.1	83.3	62.1	60.5	72.2
	CR	90.7	90.0	81.9	66.4	52.2	72.2
	FUNCODER	94.9	90.0	81.5	75.9	66.0	78.2
GPT-4o mini	Standard [†]	87.2	82.2	80.6	62.9	61.9	71.8
	CoT [†]	97.4	90.0	87.0	71.6	61.2	77.2
	PoT	89.7	81.1	76.9	63.8	61.2	71.0
	Self-Refine	87.2	82.2	80.6	62.9	61.9	71.8
	FUNCODER	94.9	82.2	81.5	62.9	63.9	73.2
Llama3 _{8b}	CoT [†]	76.9	46.7	46.3	25.9	27.9	38.6
	PoT	64.1	43.3	41.7	25.0	17.0	32.6
	FUNCODER	79.5	60.0	52.3	37.4	27.9	45.0
Codestral _{22b}	PoT	79.5	56.7	57.4	34.5	29.9	45.6
	FUNCODER	84.6	66.7	67.6	43.1	39.5	54.8
StableCode _{3b}	PoT	35.9	22.2	19.4	7.8	5.4	14.4
	FUNCODER	53.8	37.8	35.2	21.6	10.2	26.6
CodeLlama _{34b}	PoT	36.1	30.7	28.0	13.0	8.8	15.2
	FUNCODER	60.6	52.1	44.3	28.8	16.3	24.4
StarCoder2 _{15b}	PoT	43.6	44.4	45.4	20.7	14.3	30.2
	FUNCODER	71.8	57.8	55.1	26.7	23.3	40.8

B Examples

We provide example solutions for the baseline and FUNCODER on code generation and mathematical reasoning. All results are generated with GPT-3.5.

B.1 Code Generation

Info: xCodeEval/ff3c39b759a049580a6e96c66c904fdc (binary search, difficulty 1800)

Problem

```
time limit per test: 1 second
memory limit per test: 256 megabytes
input: standard input
output: standard output
```

Description

You have two friends. You want to present each of them several positive integers. You want to present cnt1 numbers to the first friend and cnt2 numbers to the second friend. Moreover, you want all presented numbers to be distinct, that also means that no number should be presented to both friends. In addition, the first friend does not like the numbers that are divisible without remainder by prime number x . The second one does not like the numbers that are divisible without remainder by prime number y . Of course, you're not going to present your friends numbers they don't like. Your task is to find such minimum number v , that you can form presents using numbers from a set $1, 2, \dots, v$. Of course you may choose not to present some numbers at all. A positive integer number greater than 1 is called prime if it has no positive divisors other than 1 and itself.

Input

The only line contains four positive integers cnt1 , cnt2 , x , y ($1 \leq \text{cnt1}, \text{cnt2} < 109$; $\text{cnt1} + \text{cnt2} \leq 109$; $2 \leq x < y \leq 3 \cdot 10^4$) - the numbers that are described in the statement. It is guaranteed that numbers x , y are prime.

Output

Print a single integer - the answer to the problem.

Sample Input 1

```
...
3 1 2 3
...
```

Sample Output 1

```
...
5
...
```

Sample Input 2

```
...
1 3 2 3
...
```

Sample Output 2

```
...
```

```

4
...

## Notes

In the first sample you give the set of numbers {1,3,5} to the first friend and the set of numbers {2} to the second friend. Note that if you give set {1, 3, 5} to the first friend, then we cannot give any of the numbers 1, 3, 5 to the second friend. In the second sample you give the set of numbers {3} to the first friend, and the set of numbers {1, 2, 4} to the second friend. Thus, the answer to the problem is 4.

```

B.1.1 Standard

```

def main() -> None:
    cnt1, cnt2, x, y = list(map(int, input().split()))
    # binary search
    l, r = 1, 2 * (cnt1 + cnt2) # upper bound
    while l < r:
        mid = (l + r) // 2
        # number of numbers that can be given to the first friend
        cnt_x = mid // x
        # number of numbers that can be given to the second friend
        cnt_y = mid // y
        # number of numbers that can be given to both friend
        cnt_xy = mid // (x * y)
        # number of numbers that can't be given to any friend
        cnt_none = mid - cnt_x - cnt_y + cnt_xy
        if cnt_x >= cnt1 and cnt_y >= cnt2 and cnt_none >= cnt1 + cnt2:
            r = mid
        else:
            l = mid + 1
    print(l)

```

Result: Wrong answer on sample test 1: expected: '5', found: '8' [1th token] ✗

B.1.2 FUNCODER

```

def main() -> None:
    cnt1, cnt2, x, y = map(int, input().split())

    left = 1
    right = int(2e9)
    while left < right:
        mid = (left + right) // 2
        if is_valid(mid, x, y, cnt1, cnt2):
            right = mid
        else:
            left = mid + 1

    print(left)

def is_valid(v, x, y, cnt1, cnt2):
    # check if there are cnt1 numbers not divisible by x
    non_x_numbers = v - (v // x)
    if non_x_numbers < cnt1:
        return False

    # check if there are cnt2 numbers not divisible by y
    non_y_numbers = v - (v // y)
    if non_y_numbers < cnt2:
        return False

```

```

# check if there are cnt1+cnt2 numbers divisible by both x and y
lcm = (x * y) // gcd(x, y)
common_numbers = v - (v // lcm)
if common_numbers < cnt1 + cnt2:
    return False

return True

def gcd(a: int, b: int) -> int:
    """get the greatest common divisor of two numbers $a$ and $b$"""

    while b:
        a, b = b, a % b
    return a

```

Result: Accepted ✓

B.2 Mathematical Reasoning

Info: MATH/number_theory/42 (level 5)

Problem What is the sum of the smallest and second-smallest positive integers a satisfying the congruence $27a \equiv 17 \pmod{40}$?

Ground Truth: 62

B.2.1 CoT

We can solve for a by multiplying both sides of the congruence by the modular inverse of 27 modulo 40 . Since 27 and 40 are relatively prime, we know that 27 has a modular inverse modulo 40 . We can find this inverse using the Euclidean algorithm:

$$\begin{aligned} \gcd(40, 27) &= \gcd(27, 13) \\ &= \gcd(13, 1) \\ &= 1. \end{aligned}$$

Therefore, the modular inverse of 27 modulo 40 exists. We can find it using the Euclidean algorithm backwards:

$$\begin{aligned} 1 &= 13 - 3 \cdot 4 \\ &= 13 - 3(40 - 27) \\ &= 4 \cdot 27 - 3 \cdot 40. \end{aligned}$$

Thus, the modular inverse of 27 modulo 40 is 4 . Multiplying both sides of the congruence $27a \equiv 17 \pmod{40}$ by 4 gives $108a \equiv 68 \pmod{40}$. Simplifying the left-hand side gives $8a \equiv 28 \pmod{40}$. Dividing both sides by 4 gives $2a \equiv 7 \pmod{10}$. The smallest positive integer a satisfying this congruence is $a=9$, and the second-smallest is $a=19$. Therefore, the sum of the smallest and second-smallest positive integers a satisfying the congruence $27a \equiv 17 \pmod{40}$ is $9+19=\boxed{28}$.

Result: 28 ✗

B.2.2 PoT

```
def solution():
    r"""What is the sum of the smallest and second-smallest positive
    integers $a$ satisfying the congruence $$27a \equiv 17 \pmod{40}~?
    $$"""
    k = 1
    candidates = []
    while len(candidates) < 2:
        a = (40 * k + 17) // 81
        if a > 0 and (40 * k + 17) % 81 == 0:
            candidates.append(a)
        k += 1

    return sum(candidates)
```

Result: 74 ✗

B.2.3 FUNCODER

```
from typing import Tuple

def solution():
    r"""What is the sum of the smallest and second-smallest positive
    integers $a$ satisfying the congruence $$27a \equiv 17 \pmod{40}~?
    $$"""

    a = 27
    m = 40
    b = 17
    inverse = find_inverse(a, m)
    x = (b * inverse) % m

    # Find the smallest and second-smallest positive integers
    # satisfying the congruence equation
    # using the formula: x + k * (m // gcd(a, m))
    gcd, _, _ = extended_euclid(a, m)
    s1 = x
    s2 = x + (m // gcd)
    return s1 + s2

def find_inverse(a: int, m: int) -> int:
    """find the inverse of $a$ modulo $m$"""

    _, x, _ = extended_euclid(a, m)
    return x % m if x >= 0 else (x + m) % m

def extended_euclid(a: int, b: int) -> Tuple[int, int, int]:
    """Extended Euclidean Algorithm"""
    if b == 0:
        return a, 1, 0
    gcd, x1, y1 = extended_euclid(b, a % b)
    x = y1
    y = x1 - (a // b) * y1
    return gcd, x, y
```

Result: 62 ✓

C Prompts

C.1 Prompt for Standard

You are a programming copilot, you can solve a problem by writing Python functions. Your task is to:

- You need to write a Python function that returns the answer.
- You can import libraries to better solve the problem.
- Do not write any code outside the function (importing is accepted)

<User>:

Let's think step by step and complete the following Python function `sum\_factor` that solves:

```
```python
def sum_factor(a: int, b: int) -> int:
 """Return the sum of all common prime factors of a and b"""

 raise NotImplementedError()
```
```

<Assistant>:

First, I need to get the prime factors of \$a\$ and \$b\$.
Second, I can use `for` loop to find common element in two factors list.
Here is the `sum\_factor` function:

```
```python
def sum_factor(a: int, b: int) -> int:
 """Return the sum of all common prime factors of a and b"""
 factors_a = prime_factor(a)
 factors_b = prime_factor(b)
 common_factors = get_common(factors_a, factors_b)
 return sum(common_factors)

def prime_factor(x: int) -> list:
 """get a list of prime factors of number x"""
 ret = []
 i = 1
 while i * i <= x:
 i += 1
 if x % i == 0 and is_prime(i):
 ret.append(i)
 return ret

def is_prime(x: int) -> bool:
 """determine x is a prime number or not"""
 if x < 2:
 return False
 for i in range(2, int(x**0.5) + 1):
 if x % i == 0:
 return False
 return True

def get_common(a: list, b: list) -> list:
 """get common element in two list a and b"""
 ret = []
 for item in a:
 if item in b:
 ret.append(item)
 return ret
```
```

```

<User>:

Let's think step by step and complete the following Python function `{
    cur_func_name}` that solves:

```python
{cur_func}
```

```

C.2 Prompt for Divide

You are a programming copilot, you can solve a problem by writing Python functions. Your task is to:

- For every turn, you need to write a Python function that returns the answer based on Current Code (not code in chat history).
- Do not modify function name, arg names, docstring in given functions.
- You can import libraries to better solve the problem.
- If a single function is too hard to solve, you can decompose it into multiple smaller functions.
- You can leave new function unimplemented for now, but write the function at the end of the code and comment what the function does.

```

<User>:

Current Code:
```python
def sum_common_factors(a: int, b: int) -> int:
 """Compute the sum of all common prime factors of a and b"""
 raise NotImplementedError()
```

```

Let's think step by step and complete the following Python function `sum\_common\_factors` that solves:

"Compute the sum of all common prime factors of \$a\$ and \$b\$"

<Assistant>:

First, I need to get the prime factors of \$a\$ and \$b\$.

Second, I can use `for` loop to find common element in two factors list.

Finally, sum the common factor list and return the answer.

Here is the `sum\_common\_factors` function:

```

```python
def sum_common_factors(a: int, b: int) -> int:
 """Compute the sum of all common prime factors of a and b"""
 factors_a = prime_factor(a)
 factors_b = prime_factor(b)
 common_factors = get_common(factors_a, factors_b)
 return sum(common_factors)

def prime_factor(x: int) -> list:
 """get a list of prime factors of number x"""
 raise NotImplementedError()

def get_common(a: list, b: list) -> list:
 """get common element in two list a and b"""
 raise NotImplementedError()
```

```

```

<User>:

Current Code:
```python
def sum_common_factors(a: int, b: int) -> int:
 """Compute the sum of all common prime factors of a and b"""
 factors_a = prime_factor(a)
 factors_b = prime_factor(b)
 common_factors = get_common(factors_a, factors_b)
 return sum(common_factors)

def get_common(a: list, b: list) -> list:
 """get common element in two list a and b"""
 raise NotImplementedError()
```

Let's think step by step and complete the following Python function `get_common` that solves:
"get common element in two list $a$ and $b$"

<Assistant>:

Here is the `get_common` function:

```python
def get_common(a: list, b: list) -> list:
 """get common element in two list a and b"""
 ret = []
 for item in a:
 if item in b:
 ret.append(item)
 return ret
```

<User>:

Current Code:


```
{prev_code}
```



Let's think step by step and complete the following Python function `{cur_func_name}` that solves:


```
{cur_func_doc}
```


```

C.3 Prompt for Conquer

You are a programming copilot, you can solve a problem by writing Python functions. Your task is to:

- For every turn, you need to write a Python function that returns the answer, based on current code (not code in chat history) and problem description.
- Do not modify function name, arg names, docstring in given functions.
- Consider reusing existing functions that are already implemented.
- You can import libraries to better solve the problem.

```

<User>:

Current Code:

```python
def prime_factor(x: int) -> list:
 """get a list of prime factors of number x"""

```

```

ret = []
i = 1
while i * i <= x:
 i += 1
 if x % i == 0 and is_prime(i):
 ret.append(i)
return ret

def is_prime(x: int) -> bool:
 """determine x is a prime number or not"""
 if x < 2:
 return False
 for i in range(2, int(x**0.5) + 1):
 if x % i == 0:
 return False
 return True

def get_common(a: list, b: list) -> list:
 """get common element in two list a and b"""
 ret = []
 for item in a:
 if item in b:
 ret.append(item)
 return ret

def sum_common_factors(a: int, b: int) -> int:
 """Return the sum of all common prime factors of a and b"""

 raise NotImplementedError()

```

Let's think step by step and implement the following method `sum\_common\_factors` using existing functions to solve:  
 "Return the sum of all common prime factors of \$a\$ and \$b\$"

<Assistant>:

First, I need to get the prime factors of \$a\$ and \$b\$.  
 Second, I can use `for` loop to find common element in two factors list.  
 Finally, sum the common factor list and return the answer.  
 Here is the `sum\_common\_factors` function:

```

```python
def sum_common_factors(a: int, b: int) -> int:
    """Compute the sum of all common prime factors of $a$ and $b$"""
    factors_a = prime_factor(a)
    factors_b = prime_factor(b)
    common_factors = get_common(factors_a, factors_b)
    return sum(common_factors)

```

<User>:

Current Code:

```

```python
{prev_code}

```

Let's think step by step and implement the following method `{cur\_func\_name}` using existing functions to solve:  
 "{cur\_func\_doc}"

## C.4 Prompt for Generate Possible Input

You are a proficient software engineer and architect, specialized in testing, capable of observing and finding security holes and logical issues in code that spans across multiple computer science fields and mathematics. You are given a piece of Python function, and are expected to produce some function calls for that function. Specifically:

- You should invoke the function in a one-liner fashion.
- Do not bring in imports other than what's already imported. Use the pre-declared imports in the original function only.
- The callee may have multiple arguments, treat them with care.
- You **must** respect the function signature and docstring, and be aware so you don't generate illegal inputs.
- Keep the inputs simple but general, and that either edge cases or common cases are meaningful.

Here follows a series of mutually uncorrelated functions to test, one per conversation.

<User>:

Let's think step by step and create some tests for the following function `check\_valid\_brackets(...)` in Python.

```
```python
def check_valid_brackets(seq: str) -> bool:
    """Determine if a bracket sequence consisting of '(', ')', '{',
    '}', '['
    and ']' is valid."""

    mapping = {')': '(', '}': '{', ']': '['}
    stack = []
    for c in seq:
        if c in mapping:
            if not stack or stack[-1] != mapping[c]:
                return False
            stack.pop()
        else:
            stack.append(c)
    return not stack
```
```

Store your function calls for `check\_valid\_brackets(...)` as function callss, one per line. They will be called later.

<Assistant>:

Sure, I can create some function calls for the `check\_valid\_brackets` function. We can either choose to test it with a valid bracket sequence or an invalid one. Empty strings are also considerable. Here are some function calls for the function:

```
```python
check_valid_brackets("()") # True
check_valid_brackets("([[]])") # True
check_valid_brackets("(()())") # False
check_valid_brackets("()[]{}") # True
check_valid_brackets("([)]") # False
check_valid_brackets("") # True
check_valid_brackets(")(") # False
```
```

<User>:

Let's think step by step and create some tests for the following function `{cur_func_name}(...)` in Python.

```
```python
{prev_code}
```
```

Store your function calls for `{cur_func_name}(...)` as function calls, one per line. They will be called later.

## C.5 Prompt for Self-Test Generation

You are a proficient software engineer and architect, specialized in testing, capable of observing and finding security holes and logical issues in code that spans across multiple computer science fields and mathematics. You are given a piece of Python function, and are expected to produce some test cases for that function. Specifically:

- You should invoke the function and assert its results in a one-liner fashion.
- Do not bring in imports other than what's already imported. Use the pre-declared imports in the original function only.
- The callee may have multiple arguments, treat them with care.
- You **must** respect the function signature and docstring, and be aware so you don't generate illegal inputs.
- Keep the inputs & outputs simple but general, and that either edge cases or common cases are meaningful.

Here follows a series of mutually uncorrelated functions to test, one per conversation.

<User>:

Let's think step by step and create some tests for the following function `lcm(...)` in Python.

```
```python
def lcm(a: int, b: int) -> int:
    """Find the least common multiple of `a` and `b`. Samples:

    >>> lcm(3, 5)
    15
    >>> lcm(4, 6)
    12
    """
    return round(a * b / gcd(a, b))
```
```

Store your test cases for `lcm(...)` as assertions, one per line. They will be called later.

<Assistant>:

Sure, I can create some test cases for the `check_valid_brackets` function. We consider the following cases: 1. the two operands are not co-prime and has common factors; 2. the two operands are equal; 3. one of them is 1; 4. two of them is 1; 5. both operands are primes. Here is an example of these test cases in Python:

```
```python
assert lcm(15, 25) == 75
assert lcm(32, 32) == 32
```
```

```

assert lcm(1, 5) == 5
assert lcm(1, 1) == 1
assert lcm(17, 19) == 17 * 19
'''

```

<User>:

Extract tests for the following function `{cur_func_name}(...)` in Python.

```

'''python
{prev_code}
'''

```

Store your test cases for `{cur_func_name}(...)` as assertions, one per line. They will be called later.

## C.6 Prompt for MATH judging

You are a mathematical teacher, your task is to:

- Judge whether the prediction is matching the answer
- Output "Judge: Correct." or "Judge: Wrong.", please do not output redundant words
- Numerical errors should be ignored ( $1$  is equal to  $0.999999998$ )
- Some answer might be represent in latex format, and some might be float number, this should be consider as correct ( $\frac{1}{2}$  is equal to  $0.5$ ,  $\sqrt{66}$  is equal to  $24.37211$ )
- Unit in answer should be ignored, and should be consider as correct ( $13 \text{ cm}^2$  is equal to  $13.0$ ,  $\$13$  is equal to  $\$13$ )

Now, the answer and prediction is:

Answer: {ground\_truth}

Prediction: {model\_output}

Please output "Judge: Correct." if two answers are literally the same, or "Judge: Wrong." for not same, please do not output redundant words.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We conduct experiments to reflect the performance of our methods in code generation (§3.1) and mathematical reasoning (§3.2), we also included analysis and ablation study in multiple aspects.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Yes, please refer to the Limitation paragraph in Discussion (§5).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include detailed information in the Appendix to support reproducibility, including method details and the model versions we use (§A.1), baseline settings (§A.2), dataset information (§A.3), evaluation metrics (§A.4), details of analysis process (§A.4), complete examples (§B) and prompts (§C).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: Our code is made available at <https://github.com/cometeme/funcoder>.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: Yes, the experimental setting and details can be found in both the main paper and the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[No\]](#)

Justification: This paper does not contain error bars or statistical significance analysis.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We include the models and compute resources in Appendix A.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: I have checked the Code Of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Yes, please refer to the Broader Impacts paragraph in Discussion (§5).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release new data or models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We reference the code (§A.2), data (§A.3) and models (§A.1) in the Appendix.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.