Top Leaderboard Ranking = Top Coding Proficiency, Always? EvoEval: Evolving Coding Benchmarks via LLM

Chunqiu Steven Xia* **Yinlin Deng*** **Lingming Zhang**

University of Illinois Urbana-Champaign

{chunqiu2, yinlind2, lingming}@illinois.edu

Abstract

Large language models (LLMs) have become the go-to choice for code generation tasks, with an exponential increase in the training, development, and usage of LLMs specifically for code generation. To evaluate the ability of LLMs on code, both academic and industry practitioners rely on popular handcrafted benchmarks. However, prior benchmarks contain only a very limited set of problems, both in quantity and variety. Further, due to popularity and age, many benchmarks are prone to data leakage where example solutions can be readily found on the web and thus potentially in training data. Such limitations inevitably lead us to inquire: *Is the leaderboard performance on existing benchmarks reliable and comprehensive enough to measure the program synthesis ability of LLMs?* To address this, we introduce EvoEval– a program synthesis benchmark suite created by *evolving* existing benchmarks into different targeted domains for a comprehensive evaluation of LLM coding abilities. Our study on **57** LLMs shows that compared to the high performance obtained on standard benchmarks like HumanEval, there is a significant drop in performance (on average 38.1%) when using EvoEval. Additionally, the decrease in performance can range from 19.6% to 47.7%, leading to drastic ranking changes amongst LLMs and showing potential overfitting of existing benchmarks. Furthermore, we showcase various insights including the brittleness of instruction-following models when encountering rewording or subtle changes as well as the importance of learning problem composition and decomposition. EvoEval not only provides comprehensive benchmarks, but can be used to further evolve arbitrary problems to keep up with advances and the ever-changing landscape of LLMs for code. We have open-sourced our benchmarks, tools, and all LLM-generated code at https://github.com/evo-eval/evoeval.

1 Introduction

Program synthesis (Gulwani et al., 2017) is regarded as the *holy-grail* in the field of computer science. Recently, large language models (LLMs) have become the default choice for program synthesis due to its code reasoning capabilities acquired through training on code repositories. Popular LLMs like GPT-4 (OpenAI, 2023), Claude-3.5 (Anthropic, 2024b), and Gemini (Team et al., 2023) have shown tremendous success in aiding developers on a wide range of coding tasks (Chen et al., 2021; Xia & Zhang, 2023; Deng et al., 2023b). Furthermore, researchers and practitioners have designed code LLMs (e.g., DeepSeek Coder (Guo et al., 2024), CodeLlama (Rozière et al., 2023), and StarCoder (Li et al., 2023)) using a variety of training methods designed specifically for the code domain to improve LLM code understanding.

Coding benchmarks like HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) have been handcrafted to evaluate the program synthesis task of turning natural language descriptions (e.g., docstrings) into code snippets. These code benchmarks measure

*Contributed equally with author ordering decided by *Nigiri*.

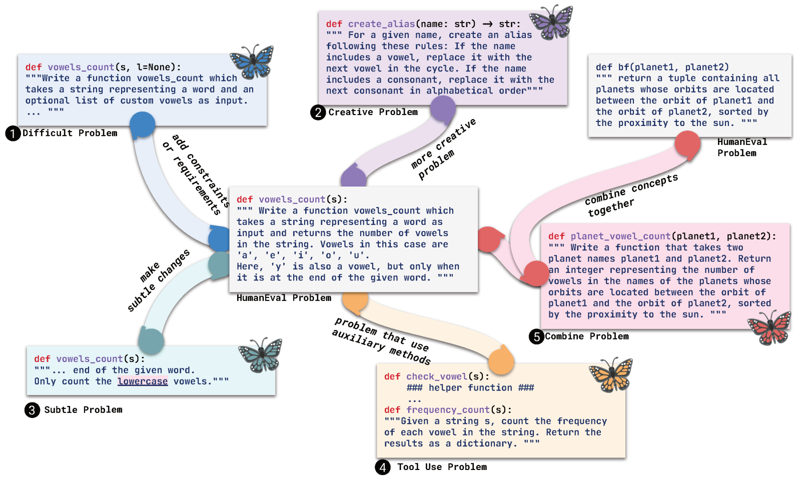

Figure 1: Exemplar problems generated in EvoEval starting from a HumanEval problem.

functional correctness by evaluating LLM-generated solutions against a set of limited predefined tests. Recent work (Liu et al., 2023) has further included augmented tests to rigorously evaluate the functional correctness of LLM generated code. However, apart from test inadequacy, existing popular code synthesis benchmarks have the following limitations:

- **Limited amount and variety of problems.** Code benchmarks are mainly constructed by human annotators manually. Due to the high manual effort required, they only contain a limited amount of problems (e.g., only 164 problems in HumanEval). Such a low amount of problems is not sufficient to fully measure the complete spectrum of program synthesis capability of LLMs. Additionally, prior benchmarks include mostly self-contained problems that lack variety in both types and domains, where the final evaluation output only shows the percentage of problems solved. While they provide a baseline overview of the coding abilities, LLM builders and users cannot gain deeper insights to exactly which problem types or coding scenarios the particular LLM may excel in or struggle with.
- **Prone to data leakage and training dataset composition.** Popular benchmarks like HumanEval and MBPP were released almost 4 years ago, with example solutions available in third-party open-source repositories. In fact, recent work (Riddell et al., 2024) has shown that there is substantial overlap between benchmark solutions and open-source training corpuses. Furthermore, the problems within these benchmarks are often simple derivatives of common coding problems. While recent LLMs have been climbing the leaderboard by achieving higher pass@1 scores (often with minimal difference between the next best model), it is unclear whether high scores achieved by LLMs are truly due to their learned coding capability or instead obtained via memorizing benchmark solutions.

As more LLMs are being constructed, trained, and used especially for code, the insufficient evaluation benchmarks raise a critical question: *Is leaderboard performance on existing benchmarks reliable and comprehensive enough to measure the program synthesis ability of LLMs?*

Our work. To address the limitation of existing benchmarks, we introduce EvoEval – a set of program synthesis benchmarks created by *evolving* existing problems. The key idea behind EvoEval is to use LLMs to automatically transform existing problems into targeted domains, enabling more comprehensive evaluations. Unlike prior benchmark construction approaches, which either obtain problems from open-source repositories (posing data leakage risks) or require manual construction (resulting in high manual effort and limited diversity), EvoEval leverages LLMs with targeted transformations to synthesize new coding problems. Specifically, we design five such transformations: *Difficult, Creative, Subtle, Combine, and Tool Use*. We then prompt GPT-4 to independently transform any existing problem in previous benchmarks into a new problem within the targeted domain.

Figure 1 shows a concrete example of EvoEval starting with an initial problem in HumanEval– `vowels_count` to count the number of vowels. 1 We first observe the transformation to a more difficult problem by asking GPT-4 to add additional requirements. This new problem contains a separate custom vowel list that makes the overall program logic more complex. ❷ We can craft a more creative problem of `create_alias` that still uses concepts like vowels and consonants but involves a much more innovative and unusual problem description. ❸ We can also make subtle changes to the problem where we only count the lowercase vowels to test if the LLM is simply memorizing the benchmark. ❹ We can additionally combine concepts from multiple problems together. In the example, we use another problem `bf` to create a new problem that returns the vowels in each planet sorted based on the orbiting order. ❺ Furthermore, we can test LLMs' ability to utilize helper functions (commonplace in real-world code repositories) to solve more complex problems. Again, we reuse the concepts of vowels from the initial problem. However, instead of directly solving the problem, the LLM can use the provided `check_vowel` helper function to simplify the solution.

Together, these transformed benchmarks are designed to introduce more challenging problems and assess different aspects of LLMs' code understanding and synthesis abilities. In EvoEval, we additionally use GPT-4 to generate the ground truth solution to each problem as well as rigorous test cases to evaluate the functional correctness of LLM-synthesized code. Finally, we manually check each generated problem and ground truth to ensure problem clarity and correctness. EvoEval serves as a way to further evolve existing benchmarks into more complex and well-suited problems for evaluation in order to keep up with the ever-growing LLM research. Our work makes the following contributions:

- **Benchmarks**: We present EvoEval– a set of program synthesis benchmarks created by evolving HumanEval problems. EvoEval includes 828 problems across 7 benchmarks, equipped with ground truth solutions and test cases to evaluate functional correctness.
- **Approach**: We propose a complete pipeline to generate new coding problems for benchmarking by evolving existing problems through targeted transformations via LLMs. Furthermore, our pipeline reduces manual checking effort by automatically refining problem inconsistencies, generating ground truth, and producing test cases.
- **Study**: We conduct a comprehensive study on **57** LLMs. We found that compared to the high performance on prior benchmarks, LLMs significantly drop in accuracy (average 38.1%) on EvoEval. Additionally, this drop is not uniform across LLMs (from 19.6% to 47.7%), leading to drastic ranking changes. We further demonstrate that certain LLMs cannot keep up their high performance when evaluated on more challenging tasks or problems in different domains, highlighting the possibility of overfitting to existing benchmarks. Moreover, we observe that instruction-following LLMs are sensitive to rephrasing or subtle changes in the problem description. They also struggle with utilizing already provided auxiliary functions. We further demonstrate that current LLMs fail to effectively compose multiple general coding concepts to solve more complex variants, or address subproblems decomposed from problems they previously solved.

2 Approach

Targeted problem transformation. We first prompt a powerful LLM to evolve an existing problem into a new one using a transformation prompt. Each transformation prompt aims to transform the existing problem in a specific manner. We define two different transformation types: **semantic-altering** – changes the semantic meaning of the problem and **semantic-preserving** – modifies the description while keeping the same semantic meaning.

Problem refinement & ground truth generation. The initial evolved problem produced by the LLM may include inconsistencies like incorrect examples. For coding benchmarks, such mistakes can lead to inaccurate evaluation. As such, we introduce a refinement pipeline to iteratively rephrase and refine the problem as needed. We first query the LLM to obtain a possible solution and test inputs for the initial problem. We then evaluate the test inputs on the solution to derive the expected outputs. Next, we instruct the LLM to refine the problem by adding or fixing the example test cases in the docstring using the computed test inputs/outputs, and then regenerate a solution. We then check if the new solution on the test inputs produces the same outputs as the previous solution. The intuition is that since the refined problem should only include minimal changes, the solution output should then remain the same in the absence of any inconsistencies. As such, if we observe differences between the two solution outputs, we ask the LLM to further revise and fix any inconsistencies and repeat the process. If both solutions agree on outputs, we return the new problem description, solution, and test cases for functional evaluation.

Manual examination & test augmentation. For each transformed problem, we carefully examine and adjust any final faults to ensure each problem and ground truth are correctly specified and implemented. We further generate additional tests using an LLM-based test augmentation technique (Liu et al., 2023). Finally, we produce EVOEVAL, a comprehensive code synthesis benchmark suite containing diverse problems to evaluate LLM coding capability across various domains. Details like transformation prompts are presented in Appendix D.

3 EVOEVAL Benchmarks & Evaluation Methodology

EVOEVAL uses HUMANEVAL problems as seeds and GPT-4 as the foundation LM to produce 828 problems across 7 different benchmarks (5 semantic-altering and 2 semantic-preserving). For the semantic-altering benchmarks, we generate 100 problems each using different seed problems from HUMANEVAL. For the semantic-preserving benchmarks, we transform all 164 problems in HUMANEVAL as we reuse the original ground truths, requiring less validation.

- **DIFFICULT**: Increase complexity by adding constraints, replacing commonly used requirements to less common ones, or introducing additional steps to the original problem.
- **CREATIVE**: Produce a more creative problem using stories or narratives.
- **SUBTLE**: Make a subtle change such as inverting or replacing a requirement.
- **COMBINE**: Combine two problems by using concepts from both problems.
- **TOOL_USE**: Produce a main problem and helper functions. Each helper function is fully implemented and provides hints or useful functionality for solving the main problem.
- **VERBOSE**: Reword the original docstring to be more verbose with descriptive language
- **CONCISE**: Reword the original docstring to be more concise using concise language.

Evaluation setup: Each LLM generated sample is executed against the test cases and evaluated using differential testing (McKeeman, 1998) – comparing against the ground truth results to measure functional correctness. We focus on greedy decoding and denote this as pass@1.

Models: We evaluate **57** LLMs (Appendix C), including both proprietary and open-source models. Further, we classify the LLMs as either base or instruction-following and discuss the effect of model variants.

Input format: To produce the code solution using each LLM, we provide a specific input prompt: For base LLMs, we let the LLM autocomplete the solution given the function header and docstring. For instruction-following LLMs, we use the recommended instruction and ask the LLM to generate a complete solution for the problem.

4 Results

4.1 LLM Synthesis & Evaluation on EVOEVAL

EVOEVAL produces more complex and challenging benchmarks for program synthesis. Table 1 shows the pass@1 performance along with the ranking of LLMs on each of the semantic-altering EVOEVAL benchmarks with the average pass@1 and ranking on all benchmarks in the last columns[1]. First, compared to the success rate on HUMANEVAL, when evaluated on EVOEVAL, all LLMs **consistently perform worse**. For example, the state-of-the-art GPT-4o, GPT-4 and Claude-3.5 models solve close to 85% of all HUMANEVAL problems but fall almost below 55% pass@1 when evaluated on the DIFFICULT problems. On average, across all benchmarks, the performance of LLMs decreased by 38.1% (DIFFICULT: 56.6%, CREATIVE: 48.2%, SUBTLE: 5.0%, COMBINE: 74.7%, and TOOL_USE: 6.1%). Additionally, this drop is not uniform across all LLMs and can range from 19.6% to 47.7%.

[1]We evaluated all 57 LLMs, however, we omitted some LLMs in Table 1 for space reasons.

Table 1: pass@1 and ranking results (* indicates tie) on the semantically-altering EvoEval and HumanEval benchmarks (including HumanEval+ scores in the parenthesis). 💬 denotes instruction-following LLMs.

	Size	HumanEval pass@1	rank	Difficult pass@1	rank	Creative pass@1	rank	Subtle pass@1	rank	Combine pass@1	rank	Tool_Use pass@1	rank	EvoEval pass@1	rank
GPT-4o💬	NA	86.0 (81.7)	1	51.0	5	64.0	2	80.0	4	51.0	*3	72.0	1	67.3	2
GPT-4-Turbo💬	NA	83.5 (80.5)	*2	50.0	*6	61.0	3	82.0	*1	45.0	5	69.0	*3	65.1	4
GPT-4💬	NA	82.3 (76.2)	*6	52.0	*2	66.0	1	76.0	6	53.0	2	68.0	*6	66.2	3
GPT-3.5-Turbo💬	NA	76.8 (69.5)	*9	33.0	*19	42.0	*13	70.0	*9	33.0	9	64.0	*10	53.1	12
Claude-3.5💬	NA	83.5 (78.0)	*2	56.0	1	60.0	4	82.0	*1	57.0	1	68.0	*6	67.8	1
Claude-3💬	NA	82.3 (75.0)	*6	50.0	*6	53.0	7	81.0	3	42.0	7	69.0	*3	62.9	7
Claude-3-haiku💬	NA	74.4 (66.5)	*13	40.0	*12	47.0	*11	65.0	*16	25.0	*12	61.0	*16	52.1	13
Claude-2💬	NA	66.5 (62.2)	*24	29.0	23	42.0	*13	64.0	*19	19.0	20	57.0	*22	46.2	21
Gemini-1.5-pro💬	NA	83.5 (76.8)	*2	52.0	*2	55.0	*5	78.0	5	43.0	6	69.0	*3	63.4	6
Gemini-1.0-pro💬	NA	62.2 (56.7)	27	37.0	*16	40.0	18	53.0	*27	23.0	*15	57.0	*22	45.4	23
PaLM-2💬	NA	40.2 (36.6)	44	18.0	*38	22.0	39	36.0	*48	3.0	*45	46.0	*35	27.5	43
DS Coder-v2-Inst💬	236b	82.9 (78.7)	5	52.0	*2	55.0	*5	75.0	7	51.0	*3	70.0	2	64.3	5
DS Coder-Inst💬	33b	78.0 (73.2)	8	47.0	9	47.0	*11	67.0	*11	31.0	*10	66.0	8	56.0	8
	6.7b	74.4 (69.5)	*13	40.0	*12	37.0	*19	61.0	*23	18.0	*21	51.0	30	46.9	20
	1.3b	63.4 (60.4)	26	20.0	*36	25.0	*31	53.0	*27	9.0	*34	39.0	*47	34.9	30
DS Coder	33b	50.6 (42.7)	32	26.0	26	23.0	*36	47.0	*32	11.0	*31	63.0	*13	36.8	29
	6.7b	45.1 (38.4)	*37	21.0	*32	24.0	*33	47.0	*32	5.0	*41	55.0	*25	32.9	35
	1.3b	29.9 (26.2)	51	6.0	*54	19.0	*41	27.0	55	0.0	57	40.0	46	20.3	51
DS Coder-1.5-Inst💬	7b	68.9 (63.4)	*21	37.0	*16	37.0	*19	66.0	*14	24.0	14	60.0	*18	48.8	16
DS Coder-1.5	7b	42.1 (34.8)	*41	21.0	*32	34.0	*23	43.0	*37	4.0	*43	54.0	*27	33.0	34
Llama -3.1-Inst💬	70b	75.0 (68.9)	*11	42.0	10	49.0	9	73.0	8	34.0	8	62.0	15	55.8	9
Llama -3-Inst💬	70b	73.8 (71.3)	*15	41.0	11	52.0	8	70.0	*9	31.0	*10	64.0	*10	55.3	10
CodeLlama-Inst💬	70b	66.5 (59.8)	*24	31.0	22	41.0	*16	65.0	*16	18.0	*21	65.0	9	47.7	18
	34b	51.8 (43.9)	31	22.0	*30	27.0	29	43.0	*37	9.0	*34	47.0	*33	33.3	33
	13b	48.8 (42.7)	35	21.0	*32	25.0	*31	46.0	35	8.0	*37	54.0	*27	33.8	32
	7b	43.3 (39.0)	39	14.0	44	18.0	*43	40.0	*43	8.0	*37	44.0	*39	27.9	42
CodeLlama	70b	60.4 (52.4)	29	25.0	27	29.0	*26	49.0	*29	14.0	*27	63.0	*13	40.1	28
	34b	52.4 (43.3)	30	15.0	43	24.0	*33	47.0	*32	11.0	*31	44.0	*39	32.2	36
	13b	42.7 (36.6)	40	18.0	*38	24.0	*33	38.0	*45	6.0	40	48.0	*31	29.4	39
	7b	39.6 (36.6)	45	10.0	*48	15.0	47	42.0	40	3.0	*45	44.0	*39	25.6	44
WizardCoder💬	34b	61.6 (54.3)	28	24.0	28	32.0	25	55.0	26	17.0	*24	55.0	*25	40.8	26
WizardCoder-1.1💬	33b	73.8 (69.5)	*15	48.0	8	48.0	10	66.0	*14	20.0	19	64.0	*10	53.3	11
XwinCoder💬	34b	68.9 (62.2)	*21	33.0	*19	42.0	*13	67.0	*11	15.0	26	60.0	*18	47.7	19
Phind-CodeLlama-2	34b	70.7 (66.5)	19	22.0	*30	35.0	22	63.0	21	25.0	*12	58.0	21	45.6	22
Code Millenials💬	34b	73.2 (69.5)	17	35.0	18	41.0	*16	65.0	*16	17.0	*24	56.0	24	47.9	17
Speechless-CL💬	34b	75.0 (69.5)	*11	38.0	15	37.0	*19	64.0	*19	23.0	*15	59.0	20	49.3	15
Magicoder-s-DS💬	6.7b	76.8 (70.7)	*9	40.0	*12	34.0	*23	67.0	*11	21.0	*17	61.0	*16	50.0	14
Magicoder-s-CL💬	7b	70.1 (65.9)	20	27.0	25	26.0	30	58.0	25	11.0	*31	52.0	29	40.7	27
StarCoder2	15b	45.1 (36.0)	*37	16.0	*41	19.0	*41	41.0	*41	5.0	*41	48.0	*31	29.0	41
	7b	34.8 (31.1)	*46	12.0	*45	17.0	45	38.0	*45	2.0	*51	46.0	*35	25.0	46
	3b	31.1 (26.2)	50	8.0	*51	14.0	*48	31.0	*50	2.0	*51	35.0	51	20.2	52
StarCoder	15b	34.8 (30.5)	*46	12.0	*45	11.0	53	37.0	47	2.0	*51	44.0	*39	23.5	47
Mixtral-Inst💬	8x7b	42.1 (38.4)	*41	21.0	*32	18.0	*43	41.0	*41	9.0	*34	45.0	*37	29.3	40
OpenChat💬	7b	71.3 (66.5)	18	33.0	*19	29.0	*26	62.0	22	14.0	*27	43.0	44	42.1	24
Gemma-Inst💬	7b	28.0 (23.2)	*54	6.0	*54	10.0	*54	29.0	54	2.0	*51	31.0	53	17.7	54
Gemma	7b	31.7 (25.0)	49	12.0	*45	13.0	*50	40.0	*43	2.0	*51	39.0	*47	23.0	48
	2b	22.0 (17.1)	57	2.0	57	6.0	57	24.0	57	2.0	*51	21.0	56	12.8	57
Phi-2	2.7b	50.0 (45.1)	*33	18.0	*38	23.0	*36	49.0	*29	14.0	*27	37.0	50	31.8	38
Qwen-1.5💬	72b	67.1 (61.6)	23	28.0	24	28.0	28	61.0	*23	21.0	*17	47.0	*33	42.0	25
	14b	50.0 (45.7)	*33	20.0	*36	23.0	*36	48.0	31	18.0	*21	44.0	*39	33.8	31
	7b	42.1 (37.8)	*41	16.0	*41	13.0	*50	43.0	*37	7.0	39	32.0	52	25.5	45

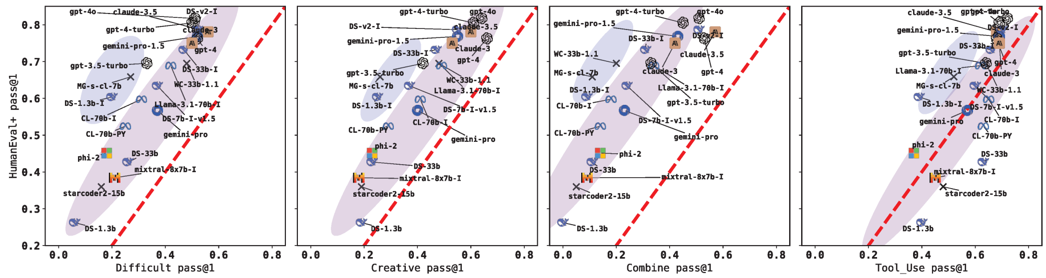

Figure 2: HUMANEVAL+ vs EVOEVAL pass@1. Red identity line shows equivalent performance. We cluster the LLMs into: purple region – aligned performance on HUMANEVAL vs. EVOEVAL and blue region – over performance on HUMANEVAL vs. EVOEVAL.

LLMs struggle on EVOEVAL benchmarks compare to the high performance achieved on HUMANEVAL. One surprising finding comes from SUBTLE, where the average performance of LLMs drops by 22.5% on the same 100 problems[2], even though only small changes are made to the original problems and the difficulty remains roughly the same. Appendix E Figure 21 presents an example problem and failing solution. Furthermore, we can also identify LLMs which struggle heavily on specific types of problems compared to their relative performance on HUMANEVAL. Figure 2 shows a scatter plot of HUMANEVAL+ vs. EVOEVAL scores. As we saw before, the significant portions of the models tend to be worse on EVOEVAL than HUMANEVAL (i.e., purple shaded region). However, there are LLMs that have a *much* higher HUMANEVAL score compared to their performance on EVOEVAL (i.e., blue shaded region). This implies potential data leakage of popular benchmarks where LLM performances are artificially inflated but do not translate to more difficult or other program synthesis problems.

Significant ranking changes of LLMs on EVOEVAL. Compared to HUMANEVAL where top models all perform similarly, we observe drastic differences in ranking changes on EVOEVAL. We observe that while the relative difference between the top 10 models on HUMANEVAL is around 10%, the difference on EVOEVAL on average is over 20%. Due to such saturation, existing benchmarks may not reliably rank the program synthesis ability of each model. For example, while Claude-3.5 and GPT-4-Turbo are tied for second on HUMANEVAL, they both excel at different types of problems: Claude-3.5 performs best on difficult and combine problems, while GPT-4-Turbo is better with tool use and creative tasks. Furthermore, while GPT-4o achieves the top HUMANEVAL and HUMANEVAL+ score, it falls off compared to the base GPT-4 variant where it is worse on DIFFICULT, CREATIVE and COMBINE problems. Such evaluation cannot be gained through naively reporting existing coding benchmark performance. Overall, by evolving the original benchmark into more difficult and diverse problems of different types, EVOEVAL can provide a more holistic evaluation and ranking of the coding ability of LLMs.

EVOEVAL can be used to comprehensively compare multiple models. In Figure 3, while both WizardCoder-1.1 and Phind-CodeLlama-2 have similar HUMANEVAL scores, they perform drastically differently across EVOEVAL benchmarks. WizardCoder-1.1 is better on DIFFICULT and CREATIVE while Phind-CodeLlama-2 is better on COMBINE problems. This can be explained through the training dataset used in each LLM: WizardCoder-1.1 uses an evolving dataset by generating more complex problems whereas Phind-CodeLlama-2 is fine-tuned on high quality programming problems that seems to boost the ability to solve programs which combines multiple programming concepts. Different from just reporting a singular pass@*k* score, EVOEVAL also allows a detailed analysis

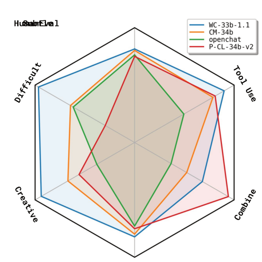

Figure 3: Radar graph

[2]Note that SUBTLE only contains 100 problems, and the pass@1 score on these 100 seed HUMANEVAL problems is higher compared to the full 164 problems. Therefore, this back-to-back performance drop is much higher than the performance drop from full HUMANEVAL to SUBTLE (5.0%) mentioned above.

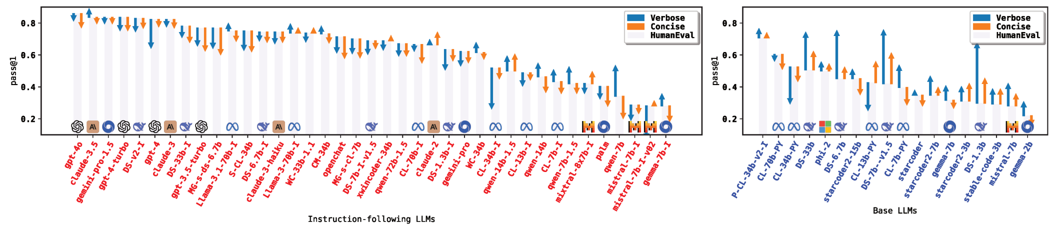

Figure 4: HUMANEVAL pass@1 with relative decrease or increase on VERBOSE and CONCISE.

Table 2: Results on COMBINE and COMBINE-NAIVE. HUMANEVAL is categorized into *pass both*, *one* and *none*, depending on the success on the two parent problems used for combination. COMBINE (Solved) and COMBINE-NAIVE (Solved) then show the distribution of solved problems that came from the previous categories. *Composition Percentage* is the % of *pass both* problems the LLM can *still* solve when combined.

	Size	HUMANEVAL			COMBINE (Solved)			Composition Percentage
		pass both	pass one	pass none	pass both	pass one	pass none	
GPT-4o	NA	80	20	0	49	2	0	61.2%
GPT-4-Turbo	NA	79	19	2	38	6	1	48.1%
GPT-4	NA	93	7	0	50	3	0	53.8%
GPT-3.5-Turbo	NA	65	34	1	24	9	0	36.9%
Claude-3.5	NA	81	18	1	49	8	0	60.5%
Claude-3	NA	81	19	0	35	7	0	43.2%
Gemini-1.5-pro	NA	86	13	1	40	3	0	46.5%
DS Coder-v2-Inst	236b	83	16	1	47	4	0	56.6%
		HUMANEVAL			COMBINE-NAIVE (Solved)			
GPT-4o	NA	881	185	8	589	50	0	66.9%
GPT-4-Turbo	NA	863	195	16	407	61	3	47.2%
GPT-4	NA	1018	55	1	768	7	0	75.4%
GPT-3.5-Turbo	NA	799	261	14	474	79	1	59.3%
Claude-3.5	NA	861	203	10	710	93	1	82.5%
Claude-3	NA	796	268	10	359	96	1	45.1%
Gemini-1.5-pro	NA	788	267	19	595	148	6	75.5%
DS Coder-v2-Inst	236b	805	252	17	598	95	5	74.3%

across different dimensions of coding capability to identify particular domains or types of coding questions an LLM struggles with or excels in.

Instruction-following LLMs are sensitive to subtle changes or rephrasing in problem docstrings. Figure 4 shows the HUMANEVAL score (bar) and the relative performance drop or improvement (arrows) on VERBOSE and CONCISE. We observe that almost all instruction-following LLMs drop in performance (average 3.4% and 4.0% decrease on VERBOSE and CONCISE respectively) when evaluated on the two semantic-preserving dataset compared to the original HUMANEVAL. This is drastically different from the base variants, where we even observe performance improvements (average 0.5% and 2.1% increase on VERBOSE and CONCISE respectively). VERBOSE and CONCISE do not change the semantic meaning of the original problem; they simply reword it in either a more verbose or concise manner. Prior work Deng et al. (2023a) has shown that by rephrasing the original problem description, one can further boost LLM performance, and we observe the similar phenomenon here mostly only for non-instruction-following models. Additionally, even on SUBTLE, where only small changes are applied, on average, instruction-following LLMs drops by 7.4% whereas base models only decrease by less than 1%. These findings across LLM types show that while instruction-tuned LLMs are expected to align better with detailed instructions, they fail to distinguish between these rephrasing or subtle changes in docstring, indicating potential memorization or contamination of prior evaluation benchmarks.

4.2 Problem Composition & Decomposition

Composing problems. The ability to compose different known concepts to solve new problems is known as *compositional generalization* (Keysers et al., 2020). This skill is essential for code synthesis, especially for complex problems in real-world programs. However, measuring compositional generalization in LLM presents a fundamental challenge since

Table 3: Results on DECOMPOSE. HUMANEVAL shows the pass/fail breakdown of the 50 seed HUMANEVAL problems. DECOMPOSE is categorized into *pass both*, *one* and *none*, based on if the LLM can solve both subproblems. *Decomp.* % and *Recomp.* % are the % of originally *passing* and *failing* problems for which the LLM can solve both subproblems respectively.

	Size	HUMANEVAL		DECOMPOSE						Decomp. %	Recomp. %
				HUMANEVAL pass			HUMANEVAL fail				
		pass	fail	pass both	pass one	pass none	pass both	pass one	pass none		
GPT-4o	NA	41	9	26	14	1	5	4	0	63.4%	55.6%
GPT-4-Turbo	NA	39	11	29	9	1	4	6	1	74.4%	36.4%
GPT-4	NA	47	3	37	10	0	0	3	0	78.7%	0.0%
GPT-3.5-Turbo	NA	33	17	19	13	1	11	4	2	57.6%	64.7%
Claude-3.5	NA	38	12	25	9	4	3	9	0	65.8%	25.0%
Claude-3	NA	39	11	26	11	2	6	5	0	66.7%	54.5%
Gemini-1.5-pro	NA	41	9	27	13	1	5	3	1	65.9%	55.6%
DS Coder-v2-Inst	236b	38	12	31	7	0	6	6	0	81.6%	50.0%

it requires controlling the relationship between training and test distributions (Shi et al., 2024). While it is not easy to control the pre-training data of LLMs, we have more control in the testing phase. Hence, we focus on program concepts that have been demonstrated to fall within the capabilities of an LLM, and explore whether this proficiency extends to the combination of program concepts. As such, we start by taking a deeper look at the COMBINE problems evolved from combining previous HUMANEVAL problems.

First half of Table 2 shows the COMBINE dataset results of the top LLMs. We observe that almost all problems solved came from the pass both category, which is intuitive as we do not expect LLMs to solve a problem composed of subproblems that it cannot already solve. However, the composition percentage is quite low, as only a few LLMs are able to achieve greater than half. This demonstrates that while state-of-the-art LLMs can achieve a high pass rate on simple programming tasks, they still struggle with composing these known concepts to address more complex problems.

Composing problems naively. Since COMBINE problems are not guaranteed to contain no additional new concepts, we build a simplified dataset for sequential composition. Let A and B be two separate problems with x as input(s) for A, we aim to create a new problem C with the same inputs where the solution can be written as $B(A(x))$. To accomplish this, the new problem combines docstrings for A and B sequentially. However, simple concatenation of docstringsleads to unclear descriptions. As such, for each problem in HUMANEVAL, we manually create two separate variants based on which order the problem may come in the new docstring. Figure 5 shows an example of how naive combination problem is constructed with the manual sequential instruction highlighted in red.

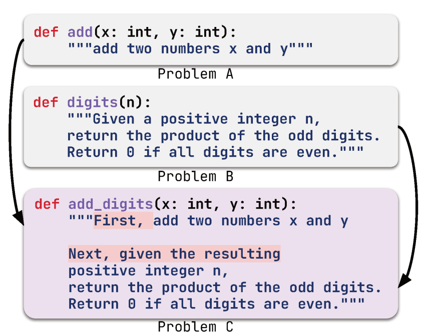

Figure 5: COMBINE-NAIVE problem

Using these modified problem docstrings, we build a sequential combination dataset – COMBINE-NAIVE, containing 1074 problems by randomly combining problems filtering for input output matching (i.e., the type of $A(x)$ should equal to the type of y in $B(y)$).

The latter half of Table 2 shows the results on COMBINE-NAIVE following the same setup as COMBINE. We observe that while the composition percentage on the naive dataset improves significantly compared to the evolved COMBINE dataset, it still fails to reach near perfection, with the best LLM being able to only solve $\sim$80% of prior pass both problems. While existing LLM training or inference paradigms for code focus on obtaining high quality datasets boosted with instruction-tuning, our result shows that existing LLMs still struggle with the concept of problem composition to tackle more complex problems.

Decomposing problems. We also evaluate *problem decomposition* – decomposing larger problems into multiple subproblems. We start by selecting 50 HUMANEVAL problems and then follow our approach in Section 2 to decompose each original problem into two smaller subproblems, creating 100 problems in our DECOMPOSE benchmark. Table 3 shows the

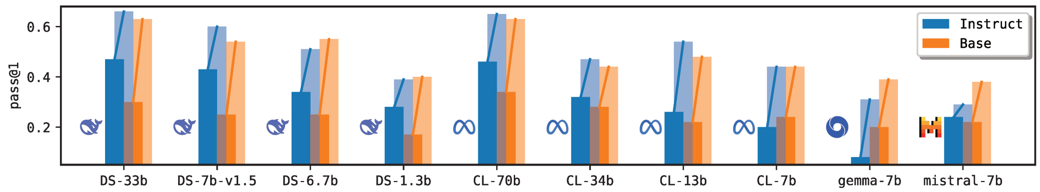

Figure 6: pass@1 from TOOL_USE-MAIN_ONLY (darker bar) to TOOL_USE (lighter bar).

results of selected LLMs on DECOMPOSE. We first observe that similar to the composition percentage in the COMBINE and COMBINE-NAIVE problems, LLMs do not achieve a high decomposition percentage. Since current LLMs are trained to recover seen outputs in their training data, and when used for program synthesis, they cannot generalize the concepts from training data. This is demonstrated by not being able to solve smaller subproblems obtained from solved more difficult parent problems. Conversely, LLMs can sometimes solve both subproblems even when the parent problem is not solved (i.e., recomposition percentage), showing room for improvement with techniques like planning (Jiang et al., 2023b) and least-to-most prompting (Zhou et al., 2022).

4.3 Tool Use Problems

We analyze the TOOL_USE benchmark, which contains helper functions. We further construct the TOOL_USE-MAIN_ONLY benchmark, which contains the same set of problem as TOOL_USE, except that the input to the LLM does not include any helpers. We observe that compared to without any helper functions (average 29.8%), LLMs on average improve by 80.1% when provided with helper functions. This is expected as helper functions reduce the work required to solve the more complex problem. However, this improvement is not uniform: the average improvement when given the auxiliary functions for instruction-following models is only 59.2% compared to the base LLMs' improvement of 122.0%.

In Figure 6, we observe that without the helpers, the instruction-following models significantly outperform their base LLMs. However, once the helpers are provided, this gap is drastically decreased, with cases even where the base models outperform their instruction-following counterparts. As real-world coding involves understanding, using, and then reusing existing functions across different places in the repository, being able to successfully leverage auxiliary methods is key. Current instruction-following LLMs are generally fine-tuned with data consisting of self-contained code snippets without the interaction and learning of function usages. This is further exacerbated by prior benchmarks, which mostly use self-contained functions, thus cannot test the tool-using capability of LLMs.

5 Related Work

Large language models for code. Starting with the general development of LLMs for general purpose tasks, developers have applied LLMs to perform code-related tasks by further training LLMs using code snippets from open-source repositories. Such LLMs include CODEX (Chen et al., 2021), CodeT5 (Wang et al., 2021), CodeGen (Nijkamp et al., 2023), InCoder (Fried et al., 2023), CodeLlama (Rozière et al., 2023), StarCoder (Li et al., 2023; Lozhkov et al., 2024), DeepSeek Coder (Guo et al., 2024), etc. More recently, researchers have applied instruction-tuning methods to train code-specific LLMs that are well-versed in following instructions. Examples of such LLMs include CodeLlama-Inst (Rozière et al., 2023) and DeepSeek Coder-Instruct (Guo et al., 2024). WizardCoder (Luo et al., 2023) instruction-tunes the model using Evol-Instruct to create more complex instructions. Magicoder (Wei et al., 2023) develops OSS-Instruct by synthesizing high quality instruction data from open-source code snippets. OpenCodeInterpreter (Zheng et al., 2024) leverages execution feedback for instruction-tuning in order to better support multi-turn code generation and refinement.

Program synthesis benchmarking. HUMANEVAL (Chen et al., 2021) and MBPP (Austin et al., 2021) are two of the most widely-used handcrafted code generation benchmarks complete with test cases. Building on these popular benchmarks, additional variants have been crafted including: EVALPLUS (Liu et al., 2023) which improves the two benchmarks with more complete test cases; HUMANEVAL-X (Zheng et al., 2023) which extends HUMANEVAL to C++, JavaScript and Go; MultiPL-E (Cassano et al., 2023) which further extends both HUMANEVAL and MBPP to 18 languages. Similarly, other benchmarks have been developed for specific domains: DS-1000 (Lai et al., 2023) and Arcade (Yin et al., 2022) for data science APIs; CodeContests (Li et al., 2022), APPS (Hendrycks et al., 2021), and LiveCodeBench (Jain et al., 2024) for programming contests, and SWE-Bench (Jimenez et al., 2024) for software engineering tasks. Different from prior benchmarks which require handcraft problems from scratch – high manual effort or scrape open-source repositories or coding contest websites – leading to unavoidable data leakage, EVOEVAL directly uses LLMs to *evolve* existing benchmark problems to create new complex evaluation problems. Furthermore, contrasting with the narrow scope of prior benchmarks (often focusing on a single type or problem, i.e., coding contests), EVOEVAL utilizes targeted transformation to evolve problems into different domains, allowing for a more holistic evaluation of program synthesis using LLMs.

6 Conclusion

We present EVOEVAL– a set of program synthesis benchmarks created by *evolving* existing problems into different target domains for a holistic and comprehensive evaluation of LLM program synthesis ability. Our results on **57** LLMs show drastic drops in performance (average 38.1%) when evaluated on EVOEVAL. Additionally, we observe significant ranking differences compared to prior leaderboards, indicating potential dataset overfitting on existing benchmarks. We provide additional insights, including the brittleness of instruction-following LLMs as well as the limited compositional generalization abilities of LLMs.

7 Acknowledgment

We thank Owen Colegrove for his help on starting this project and providing valuable feedback throughout, Jiawei Liu for providing helpful discussions and Yifeng Ding for his help in running experiments. This work was partially supported by NSF grant CCF-2131943 and Kwai Inc. This project is supported, in part, by funding from Two Sigma Investments, LP. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Two Sigma Investments, LP.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Anthropic. Introducing claude 2.1. `https://www.anthropic.com/news/claude-2-1/`, 2023.

Anthropic. Introducing the next generation of claude. `https://www.anthropic.com/news/claude-3-family/`, 2024a.

Anthropic. Introducing claude 3.5 sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet/`, 2024b.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Xiaodong Deng Kai Dang, Yang Fan, Wenbin Ge, Fei Huang, Binyuan Hui, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Tianyu Liu, Keming Lu, Jianxin Ma, Rui Men, Na Ni, Xingzhang Ren, Xuancheng Ren, Zhou San, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Jin Xu, An Yang, Jian Yang, Kexin Yang, Shusheng Yang, Yang Yao, Jianwei Zhang Bowen Yu, Yichang Zhang, Zhenru Zhang, Bo Zheng, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Introducing qwen1.5. https://qwenlm.github.io/blog/qwen1.5/, 2023b.

BudEcosystem. Code millenials 34b. URL https://huggingface.co/budecosystem/code-millenials-34b.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, et al. Multipl-e: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering*, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. Rephrase and respond: Let large language models ask better questions for themselves, 2023a.

Yinlin Deng, Chunqiu Steven Xia, Haoran Peng, Chenyuan Yang, and Lingming Zhang. Large language models are zero-shot fuzzers: Fuzzing deep-learning libraries via large language models. In *32nd International Symposium on Software Testing and Analysis (ISSTA)*, 2023b.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Scott Yih, Luke Zettlemoyer, and Mike Lewis. Incoder: A generative model for code infilling and synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=hQwb-lbM6EL.

Google. Our next-generation model: Gemini 1.5. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/, 2024.

Sumit Gulwani, Oleksandr Polozov, and Rishabh Singh. Program synthesis. *Foundations and Trends® in Programming Languages*, 4(1-2):1–119, 2017. ISSN 2325-1107. doi: 10.1561/2500000010. URL http://dx.doi.org/10.1561/2500000010.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

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. *NeurIPS*, 2021.

Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

HuggingFace. Hugging face, 2022. https://huggingface.co.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint*, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.

Xue Jiang, Yihong Dong, Lecheng Wang, Qiwei Shang, and Ge Li. Self-planning code generation with large language model. *arXiv preprint arXiv:2303.06689*, 2023b.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygcCnNKwr.

Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-tau Yih, Daniel Fried, Sida Wang, and Tao Yu. Ds-1000: A natural and reliable benchmark for data science code generation. In *International Conference on Machine Learning*, pp. 18319–18345. PMLR, 2023.

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

Yujia Li, David 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. *Science*, 2022. URL https://www.science.org/doi/abs/10.1126/science.abq1158.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1qvx610Cu7.

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 McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder 2 and the stack v2: The next generation, 2024.

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

William M McKeeman. Differential testing for software. *Digital Technical Journal*, 10(1): 100–107, 1998.

Meta. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/, 2024.

Microsoft Research. Phi-2: The surprising power of small language models. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/, 2023.

Mistral AI team. Mixtral of experts a high quality sparse mixture-of-experts. https://mistral.ai/news/mixtral-of-experts/, 2023.

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*, 2023. URL https://openreview.net/forum?id=iaYcJKpY2B_.

OpenAI. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, 2022.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

phind team. Beating gpt-4 on humaneval with a fine-tuned codellama-34b. https://www.phind.com/blog/code-llama-beats-gpt4, 2023.

Nikhil Pinnaparaju, Reshinth Adithyan, Duy Phung, Jonathan Tow, James Baicoianu, and Nathan Cooper. Stable code 3b. URL https://huggingface.co/stabilityai/stable-code-3b.

Martin Riddell, Ansong Ni, and Arman Cohan. Quantifying contamination in evaluating code generation capabilities of language models. *arXiv preprint arXiv:2403.04811*, 2024.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Kensen Shi, Joey Hong, Yinlin Deng, Pengcheng Yin, Manzil Zaheer, and Charles Sutton. Exedec: Execution decomposition for compositional generalization in neural program synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oTRwljRgiv.

Jiangwen Su. Code millenials 34b. URL https://huggingface.co/uukuguy/speechless-codellama-34b-v2.0.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Xwin-LM Team. Xwin-lm. https://github.com/Xwin-LM/Xwin-LM, 2023.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*, 2023.

Yue Wang, Weishi Wang, Shafiq Joty, and Steven CH Hoi. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8696–8708, 2021.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Source code is all you need. *arXiv preprint arXiv:2312.02120*, 2023.

Chunqiu Steven Xia and Lingming Zhang. Keep the conversation going: Fixing 162 out of 337 bugs for $0.42 each using chatgpt. *arXiv preprint arXiv:2304.00385*, 2023.

Pengcheng Yin, Wen-Ding Li, Kefan Xiao, Abhishek Rao, Yeming Wen, Kensen Shi, Joshua Howland, Paige Bailey, Michele Catasta, Henryk Michalewski, Alex Polozov, and Charles Sutton. Natural language to code generation in interactive data science notebooks. 2022.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, et al. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*, 2023.

Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhu Chen, and Xiang Yue. Opencodeinterpreter: Integrating code generation with execution and refinement. *arXiv preprint arXiv:2402.14658*, 2024.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

Qihao Zhu, Daya Guo, Zhihong Shao, Dejian Yang, Peiyi Wang, Runxin Xu, Y Wu, Yukun Li, Huazuo Gao, Shirong Ma, et al. Deepseek-coder-v2: Breaking the barrier of closed-source models in code intelligence. *arXiv preprint arXiv:2406.11931*, 2024.