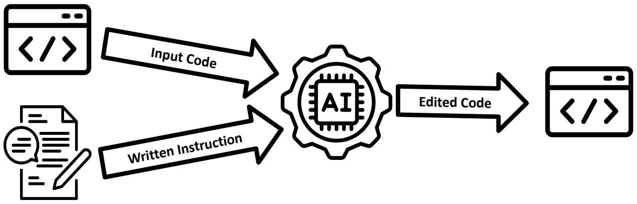
Can It Edit? Evaluating the Ability of Large Language Models to Follow Code Editing Instructions

Federico Cassano, Luisa Li, Akul Sethi, Noah Shinn, Abby Brennan-Jones, Jacob Ginesin, Edward Berman, George Chakhnashvili, Anton Lozhkov, Carolyn Jane Anderson, Arjun Guha

Instructional Code Editing with LLMs:

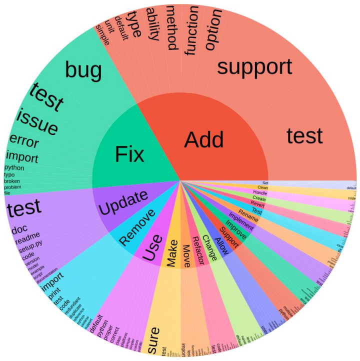


Can It Edit is a hand-crafted benchmark for code editing:

- 105 handwritten Python problems
- Each problem contains both a lazy and a descriptive instruction to complete the task
- Passing criteria determined by ground truth test execution
- Taxonomized, split between change kinds:
  a. Adaptive: introduces new functionality
  b. Perfective: enhances existing features
  c. Corrective: fixes incorrect code

Includes two training splits derived from filtering GitHub commits, for improving LLMs's code editing abilities

Dataset Statistics		
	EditPackFT	**Commits2023FT**
Total Commits	22,602	24,129
Unique Initial Verbs	184	199
Code Segments (Mean ± Std. Dev.)		
Lines of Code	29.2 ± 13.7	119.3 ± 75.9
Levenshtein Distance	197.1 ± 260.6	406.6 ± 631.2
Commit Messages (Mean ± Std. Dev.)		
Tokens	10.1 ± 4.6	23.1 ± 35.2



Evaluation Results

Model		Descriptive		Lazy	
Name	Size	*pass@1*	*ExcessCode*	*pass@1*	*ExcessCode*
Closed Models					
GPT-4	—	**63.33**	0.15 ± 0.09	**51.95**	0.14 ± 0.10
GPT-3.5-Turbo	—	48.14	0.47 ± 0.34	42.71	0.00 ± 0.00
Open Models					
CodeLlama-Instruct	70b	45.05	0.28 ± 0.15	37.52	0.02 ± 0.02
Mixtral-Instruct	8x7b	30.10	0.40 ± 0.16	24.90	0.01 ± 0.01
EditCoder	33b	**55.90**	0.33 ± 0.21	**42.33**	0.27 ± 0.24
DeepSeekCoder-Instruct	33b	49.78	0.36 ± 0.24	38.94	0.51 ± 0.34
DeepSeekCoder-Base	33b	47.71	0.53 ± 0.24	34.71	0.62 ± 0.41
CodeLlama-Instruct	34b	30.63	0.33 ± 0.21	24.15	0.18 ± 0.14
StarCoder2	15b	41.95	0.36 ± 0.20	31.48	0.04 ± 0.04
StarCoder	15b	37.10	0.56 ± 0.28	27.62	0.42 ± 0.34
OctoCoder	15b	34.43	0.12 ± 0.07	25.95	0.07 ± 0.07
CodeLlama-Instruct	13b	26.90	0.90 ± 0.68	16.89	0.42 ± 0.41
EditCoder	6.7b	48.33	0.36 ± 0.17	39.29	0.32 ± 0.25
DeepSeekCoder-Instruct	6.7b	41.03	0.13 ± 0.06	31.65	0.22 ± 0.12
DeepSeekCoder-Base	6.7b	32.62	1.01 ± 0.42	27.76	1.25 ± 0.98
CodeLlama-Instruct	7b	32.83	0.31 ± 0.15	23.49	0.36 ± 0.26
EditCoder	1.3b	26.67	0.14 ± 0.09	21.43	0.20 ± 0.12
DeepSeekCoder-Instruct	1.3b	26.22	0.32 ± 0.18	17.27	0.32 ± 0.13
DeepSeekCoder-Base	1.3b	17.90	0.69 ± 0.42	11.76	2.79 ± 2.29

Model		Corrective		Adaptive		Perfective	
Name	Size	*p@1*	*ExcessCode*	*p@1*	*ExcessCode*	*p@1*	*ExcessCode*
GPT-4	—	62.21	0.05 ± 0.03	57.29	0.31 ± 0.19	53.43	0.08 ± 0.06
GPT-3.5-Turbo	—	47.93	0.00 ± 0.00	42.29	0.17 ± 0.12	46.07	0.60 ± 0.54
EditCoder	33b	56.86	0.02 ± 0.02	51.21	0.77 ± 0.42	39.29	0.05 ± 0.04
EditCoder	6.7b	48.64	0.00 ± 0.00	42.71	0.43 ± 0.21	40.07	0.66 ± 0.42
EditCoder	1.3b	26.36	0.11 ± 0.10	23.21	0.14 ± 0.10	22.57	0.26 ± 0.18

Evaluation Details

- For each problem, we sample 20 completions at temperature 0.2, for both the lazy and descriptive instruction separately
- We calculate the pass@1 metric based on ground truth test execution of the edited code

Findings

- Closed source models outperform open ones
- Descriptive instructions yield better performance than lazy ones, despite holding same information
- EditCoder, a DeepSeek fine-tune on the commit splits, outperforms all other models of its size
- Using code coverage, we track unexecuted code generated by the models, and find that smaller models tend to generate more superfluous code.