Rethinking Code Refinement: Learning to Judge Code Efficiency

Minju Seo[1] **Jinheon Baek**[1] **Sung Ju Hwang**[1,2]
KAIST[1], DeepAuto[2]
{minjuseo, jinheon.baek, sjhwang82}@kaist.ac.kr

Abstract

Large Language Models (LLMs) have demonstrated impressive capabilities in understanding and generating codes. Due to these capabilities, many recent methods are proposed to automatically refine the codes with LLMs. However, we should rethink that the refined codes (from LLMs and even humans) are not always more efficient than their original versions. On the other hand, running two different versions of codes and comparing them every time is not ideal and time-consuming. Therefore, in this work, we propose a novel method based on the code language model that is trained to judge the efficiency between two different codes (generated across humans and machines) by either classifying the superior one or predicting the relative improvement. We validate our method on multiple programming languages with multiple refinement steps, demonstrating that the proposed method can effectively distinguish between more and less efficient versions of code.

1 Introduction

Large Language Models (LLMs) have shown significant success across a wide range of tasks, extending from natural language understanding to programming-related activities (Brown et al., 2020; Chen et al., 2021; Achiam et al., 2023; Touvron et al., 2023; Rozière et al., 2023; Abdin et al., 2024). Specifically, thanks to their capabilities in understanding and generating codes, LLMs are able to allow developers to save time, reduce errors, and boost their productivity (Shen et al., 2022). For example, several recent studies have utilized LLMs to optimize and refine the existing code bases (Madaan et al., 2023; Chen et al., 2023b). Also, more recent work has been proposed to iteratively refine the generated codes from LLMs by judging them with LLMs (Zelikman et al., 2023). Another noteworthy approach involves using LLMs to check the functional correctness of the generated codes from LLMs (Dong et al., 2023a).

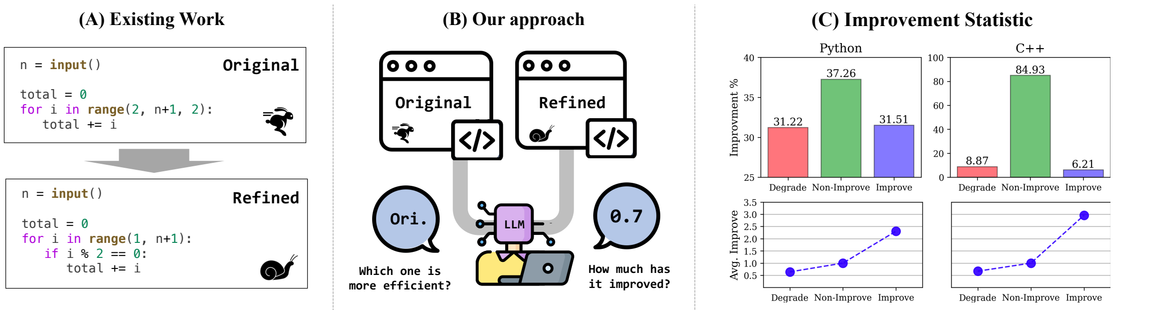
However, despite huge advancements made in the field of LLMs for code generation, the aforementioned studies assume that the codes generated and refined from LLMs are more efficient than their originals. However, as shown in Figure 1, our observations contradict this assumption, showing that LLM-generated and -refined codes do not always perform better. To handle this issue, while one may calculate the efficiencies of codes both before after modifications by actually executing them, this process introduces unnecessary inefficiencies, and may be very costly and time-consuming.

In this work, to overcome those challenges, we introduce a new task of judging the efficiency of the refined code over its original version. In addition, not only LLMs but also human coders may degrade the efficiency of codes during refinement; thus, our task of judging the efficiency between two codes involves all the possible pairs of code modification sources, including human-human, human-machine, and machine-machine. Then, to address this new task, we propose a new model (based on the code LM) that is trained to compare efficiencies between two codes. Specifically, given a code pair (one is original and the other is refined from it), the code LM is trained to classify which one is more efficient or predict how much the refined code is efficient.

We experimentally validate the effectiveness of our efficiency judgement model on multiple code refinement scenarios with multiple programming languages. The results show that our judgement model can identify the more efficient code among two different versions, substantially outperforming baselines. Our contributions are as follows:

- We point out that the refined codes from LLMs or humans do not always improve their efficiency.
- We introduce a novel approach that judges the efficiency of two different versions of codes.
- We validate our model on multiple code refinement scenarios, demonstrating its effectiveness.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 11045–11056
November 12-16, 2024 ©2024 Association for Computational Linguistics



```
n = input()

total = 0
for i in range(2, n+1, 2):
   total += i
```

```
n = input()

total = 0
for i in range(1, n+1):
   if i % 2 == 0:
      total += i
```

Figure 1: (A) Existing code refinement approaches sometimes generate the code which has inferior efficiency to the original code. (B) Our proposed approach identifies the efficient code among two different versions of codes (before and after modifications), and further predicts its relative improvement. (C) We categorize the refined code according to its efficiency gain (%) compared to the original into three classes: Degradation (less than 0.9), Non-Improvement (0.9 to 1.1), and Improvement (greater than 1.1).

2 Related Work

Large Language Models for Code Large Language Models (LLMs), trained on extensive corpora with multi-billion parameters, exhibit remarkable performance across a broad spectrum of tasks involving both text and code (Brown et al., 2020; Li et al., 2022; Rozière et al., 2023; Li et al., 2023; Achiam et al., 2023; Abdin et al., 2024; Guo et al., 2024). These models, particularly those trained on code-specific datasets, have opened up a new era in software development by not only assisting with basic programming tasks but also enabling more complex activities such as code generation (Chen et al., 2023a; Nijkamp et al., 2023), translation (Yin and Neubig, 2018; Rozière et al., 2020; Cassano et al., 2023), and refinement (Yu et al., 2023; Shirafuji et al., 2023). As such, these models are increasingly integrated into development environments, optimizing workflows, and reducing the time for development (Shen et al., 2022; Dong et al., 2023b).

LLMs for Code Refinement Beyond basic code generation, LLMs are widely used to refine the existing code bases. One of the early work in code refinement aims to detect and fix bugs in the codes by pre-training a transformer-based model on English and source code, and then fine-tuning it on commits (relevant to the part fixing bugs and improving performance) (Garg et al., 2022). In a similar vein, another work proposes to refine the code with a sequence-to-sequence model that is trained to transform the original code to its optimized version of the code (Chen et al., 2023b). Additionally, recent work showcases that LLMs are able to recursively improve their own generated codes, progressively enhancing their outputs (Zelikman et al., 2023). Further, Madaan et al. (2023) demonstrate that LLMs with sophisticated prompting strategies (to adapt LLM for code optimization) can surpass human-level performance in code optimization tasks. However, despite these substantial achievements, prior studies have primarily focused on code enhancement with limited attention to the actual efficiency of the refined code. Meanwhile, we focus on a different angle, proposing to judge the efficiency of the refined codes in advance (without executing them) based on our observation that not all the refined codes have better efficiency.

LLM-Powered Code Evaluation The objective of our work which aims to evaluate the efficiency between two codes based on LLMs has a similarity to work on LLMs for code evaluation. Early work on it uses either a term-matching-based approach (similar to BLEU) or an embedding-based approach (whose representations are obtained from language models), to compare two codes (Ren et al., 2020; Zhou et al., 2023). However, as collecting ground-truth answers for every evaluation is difficult, recent work has shifted towards using LLMs to judge the quality of the generated code, such as its utility or functional correctness, without the need for comparisons to the reference code (Dong et al., 2023a; Zhuo, 2024). Yet, unlike these approaches evaluating the single instance of the code other than the efficiency, we aim to compare efficiency in the setting where the code pair is given.

3 Method

We first provide a general description of code refinement, and then introduce our approach.

3.1 Code Refinement

Let us assume that the existing code base is defined as $\boldsymbol{c}$, which consists of a sequence of tokens as follows: $\boldsymbol{c} = \{c_1, ..., c_n\}$. Then, the objective of code refinement (in this work) is to transform the existing code base $\boldsymbol{c}$ into its improved version $\boldsymbol{c}' = \{c'_1, ..., c'_m\}$, where the execution time for $\boldsymbol{c}'$ should be faster than $\boldsymbol{c}$, formalized as follows:

Table 1: Main results on the task of judging the code efficiency, where easy denotes the dataset containing only the code pairs whose efficiency difference is more than 10%.

	All			**Easy**		
	Python	C++	All	Python	C++	All
Zero-shot	50.87	45.30	46.55	47.04	51.21	49.70
Few-shot	51.21	51.17	51.18	49.16	48.22	48.56
Zero-shot CoT	50.35	51.69	51.39	48.94	52.10	50.95
Few-shot CoT	50.52	50.87	50.79	49.50	48.09	48.60
GPT-3.5	54.50	52.17	52.69	54.97	55.59	55.37
GPT-4o	63.67	56.03	57.75	66.03	60.48	62.49
Ours	**72.49**	**62.08**	**64.42**	**77.65**	**70.14**	**72.86**

$\mathsf{Exec}(\boldsymbol{c}) > \mathsf{Exec}(\boldsymbol{c}')$. Here, Exec is the code execution function that returns its runtime.

It is worth noting that, in this work, we consider three different scenarios of code refinement. The first scenario involves a human-human interaction, where one developer revises the code originally authored by another. The second scenario, termed the human-machine scenario, consists of collaborative efforts between humans and machines. This practice has become increasingly prevalent in real-world software development environments, thanks to CodeLLMs (Chen et al., 2021; Rozière et al., 2023). Lastly, the machine-machine scenario involves autonomous code refinement by machines, a process that has shown promise in various studies (Zelikman et al., 2023; Madaan et al., 2023).

Note that, despite the huge advancements made in the field of code refinement, we find that modified codes from machines can occasionally reduce the efficiency of the original codes. Similarly, human developers may diminish the efficiency of the codes during refinement. On the other hand, executing the pair of original and modified codes at every refinement step is inefficient and time-consuming.

3.2 Judging Code Efficiency

To overcome the aforementioned limitation, we aim to predict the efficiency of the modified code over its original code, without actually executing them. This can be formulated by either the classification problem (where we classify the superior code) or the regression problem (where we predict the relative improvement of the modified code over the original code), given a pair of original and modified codes. Also, note that we operationalize classification and regression problems with CodeLLMs, due to their capabilities in understanding codes.

Specifically, given the code pair (e.g., $\boldsymbol{c}$ and $\boldsymbol{c}'$), we concatenate it and provide it with the CodeLLM, formalized as follows: $o = \mathsf{CodeLLM}([\boldsymbol{c}, \boldsymbol{c}'])$ where $[\cdot]$ is the concatenation operation, and o is the prediction output. Then, for the classification problem, we formulate it as the next token prediction task: $L_C = -\log p(o' | [\boldsymbol{c}, \boldsymbol{c}'], o' \in \{A, B\})$, where o' is ground truth that belongs to one of A or B, in which A represents the improvement over the original code and vice versa for B. Similarly, for the regression problem, we train the model to predict the relative improvement of the refined code over its original code by minimizing this prediction value with the actual relative improvement.

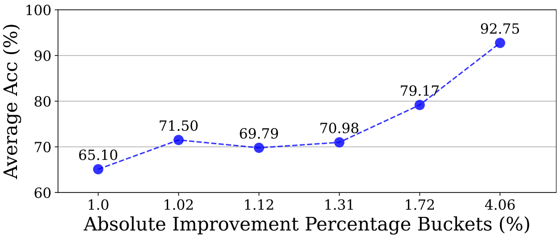


Figure 2: Results with bucketing the code pairs according to their absolute relative improvement in efficiency, on Python.

4 Experiment

We now describe experimental setups and results. We provide our code at https://github.com/going-doer/judge_code_efficiency, for reproducibility.

4.1 Experimental Setups

Dataset To validate the efficacy of our approach to judge code efficiency, we should collect pairs of two different versions of codes before and after modifications. Here, we consider three different scenarios of code editing, and, for the cases where humans refine the code, we use a dataset of code edits made by humans from Madaan et al. (2023). For the other scenarios where the machine improves the human- or machine-generated codes, we prompt the Code LLM (namely DeepSeek-Coder-Instruct-7B) (Guo et al., 2024) to refine the given codes for better efficiency. Specifically, starting with the codes generated by humans from the existing dataset, we generate the machine-refined codes with the Code LLM. In addition, from those machine-generated codes, we similarly prompt the Code LLM to improve them. Through these steps, we can obtain pairs of human-human, human-machine, and machine-machine code versions.

Baselines and Our Model In this work, as we tackle a novel problem of judging code efficiency, there are no direct baselines available to compare. Therefore, we turn to compare our approach against the basic models powered by LLMs. Specifically, given a code pair, we perform zero-shot and few-

Table 2: Breakdown results for varying the code refinement scenarios. ’H’ indicates Human and ’M’ indicates Machine.

	Statistics			**Breakdown Acc**		
Scenarios	Avg. Improve	Degrade %	Improve %	Python	C++	All
H-H	1.08	37.19	21.85	80.43	64.43	67.88
M-M	1.32	30.69	32.25	60.77	55.84	57.03
H-M	1.29	31.26	30.76	67.27	61.55	62.87

Table 3: Results with predicting the relative difference of the modified code over its original code in efficiency. Corr denotes the Spearman’s rank correlation coefficient, and Acc denotes the accuracy where we convert prediction values into classes.

	All		**Easy**	
	Corr	Acc	Corr	Acc
Python	0.66	76.38	0.64	82.88
C++	0.50	66.69	0.63	80.16
Python & C++	0.56	68.87	0.66	81.17

Table 4: Generalization results by varying the training data.

Training Datasets	Python	C++	All
Python	**72.75**	58.26	61.52
C++	57.53	60.32	57.90
Python & C++	72.49	**62.08**	**64.42**

Table 5: The performances of baselines and our model (trained on Python and C++) to other low-resource programming languages (such as Ruby, Perl, and Rust).

	All				**Easy**			
	Ruby	Perl	Rust	All	Ruby	Perl	Rust	All
Zero-shot	52.30	40.00	49.13	51.08	50.00	45.83	46.67	49.22
Few-shot	52.06	40.00	49.57	50.99	49.01	45.83	46.67	48.44
Zero-shot CoT	52.30	42.00	48.70	51.08	53.96	45.83	33.33	50.78
Few-shot CoT	52.06	40.00	49.57	50.99	49.01	45.83	46.67	48.44
Ours	**58.47**	**70.00**	**55.22**	**58.32**	**71.78**	**66.67**	**60.00**	**69.92**

shot prompting with LLMs, to decide which one is more efficient. In addition, we also enhance those strategies with Chain-of-Thought prompting (Wei et al., 2022), to elicit the reasoning ability of LLMs with the instruction: "Let’s think step by step". For our model, we use the classifier (predicting the class of the efficient code) for main experiments, and provide the performance of the regression model (predicting the relative improvement) during analysis. We use DeepSeek-Coder-Instruct-1.3B for all models, and also provide the results with GPT-3.5, GPT-4o for benchmarking the performance of stronger LLMs without finetuning.

4.2 Experimental Results

Main Results We report the main results in Table 1, and, from this, we observe that our method consistently outperforms all baseline models across all settings. Surprisingly, we find that our approach is substantially superior to GPT-3.5 and GPT-4o, demonstrating the continued limitations of even larger models in judging the code efficiency, which further supports the efficacy of our training strategy for it. In addition, our model is particularly effective in scenarios where there is a clear difference in code efficiency — specifically, a difference exceeding 10% (the easy setting). To examine the performance of our model more granularly based on varying degrees of efficiency differences between code pairs, we bucketize the code pairs based on their efficiency differences. As shown in Figure 2, the performance of our model increases when the difference between two codes becomes larger.

Results with Varying Refinement Scenarios It is worth noting that our code refinement scenario is categorized as human-human, human-machine, and machine-machine, and we report their breakdown results in Table 2. From this, we first observe that the percentage of efficiency improvement in code refinement scenarios is low, which is around 20% to 30%, and it is similar to the percentage of efficiency degradation. In addition, the average improvement made by machines is 30%, meanwhile, the improvement by humans is around 10%. On the other hand, as shown in the Breakdown Acc column, our model can more effectively identify the code improved by humans (rather than machines).

Relative Improvement Prediction Results In addition to classifying the efficient code given the code pair, we can further predict how much it is improved. For this task, we measure the performance of our model, by ranking all the code pairs based on their relative improvements and comparing them with predicted improvements. Also, if the prediction score exceeds 1.00, we classify this case that there is an improvement during refinement. As shown in Table 3, we observe both the high rank correlation coefficient and high accuracy, demonstrating the effectiveness of our approach even in this actual improvement prediction setting.

Generalization on Different Languages To see the generalization ability of our approach to different programming languages, we train the model with Python, C++ or both, and measure the performance on Python and C++ as well as relatively low-resource programming languages (such as Ruby, Perl, and Rust). As shown in Table 4 and Table 5, we observe that the model trained on one language can be generalizable to the other language, perhaps due to the algorithmic similarities in their codes. Also, this results confirm the broader applicability of our approach to various programming languages.

Table 6: With different Code LLMs, we report their average relative improvement (Avg), as well as the percentage of their degradation and improvement in efficiency. Acc (H) and (M) denote the accuracy on human- and machine-generated codes.

LLM	Avg	Degrade %	Improve %	Acc (H)	Acc (M)
DSC	1.30	31.22	31.51	80.43	65.49
CodeQwen	1.33	28.01	29.12	80.60	58.86
Granite	1.13	21.46	24.55	76.69	48.97
GPT-3.5	1.00	25.34	21.45	78.29	53.72

Results with Different Code LLMs To see the performance of different Code LLMs in refining the given codes (in terms of efficiency), and to see the performance of our efficiency judgement model trained with the code pairs constructed by different LLMs, we change the code refinement model from DeepSeekCoder (DSC) to recent CodeQwen (Bai et al., 2023), Granite (Mishra et al., 2024), and GPT-3.5. As shown in Table 6, we find that DSC and CodeQwen are superior in improving the efficiency of codes when refining them. Yet, the percentage of improvement made by each model is comparable to the percentage of degradation, which supports again our motivation that we should rethink code refinement. Lastly, our model is able to identify the more efficient code among two different versions of models made across different LLMs.

Results with Larger Models We conduct an auxiliary analysis to see how the performance of different methods changes if a model larger than DeepSeek-Coder-Instruct-1.3B (that we use for main experiments) is used. Specifically, we use its 7B model as the base code LLM and then classify the efficient code given code pairs. As shown in Table 7, we observe results similar to those obtained from the smaller 1.3B model, where our model is consistently superior to other baselines.

Visualization of Rank Correlation To visualize how accurate the predicted results of relative improvements of code pairs from our model are, we compare their ranks with the ground-truth ranks calculated by actual relative improvements of code pairs. As shown in Figure 3 where we present a scatter plot of rank correlations along with their coefficient value, we observe that the results from our approach have a positive correlation with the ground truth, demonstrating its effectiveness in predicting the relative improvement of code pairs.

5 Conclusion

In this work, we pointed out that the codes refined by humans or machines are sometimes inferior than originals, and to tackle this, we introduced a novel approach to identify the more efficient code given a pair of codes before and after modifications. We validated our method on multiple code editing scenarios involving both humans and machines, showcasing its substantial efficacy despite its simplicity.

Table 7: Results on the DeepSeek-Coder-Instruct-7B model.

	All			Easy		
	Python	C++	All	Python	C++	All
Zero-shot	52.50	47.91	48.88	51.62	49.05	49.98
Few-shot	54.50	49.86	50.90	55.87	49.81	52.00
Zero-shot CoT	54.50	49.86	50.90	55.53	52.73	53.75
Few-shot CoT	53.11	49.99	50.69	47.04	52.16	50.3
Ours	**73.44**	**62.90**	**65.27**	**78.88**	**72.68**	**74.93**

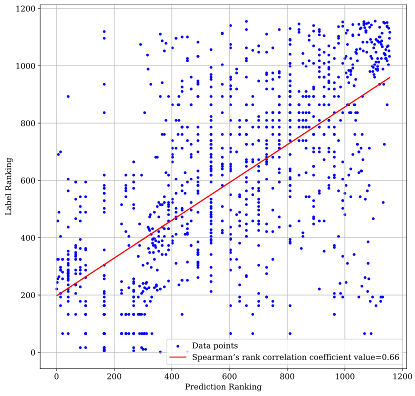


Figure 3: Visualization of the Spearman's rank correlation between the ranks of the actual relative improvements and the predicted relative improvements of code pairs, for our model.

Limitation

There are some areas that future work may improve upon. First, we perform experiments with two widely used programming languages, such as Python and C++, and it may be promising to consider other languages, particularly those used less frequently. In addition, in terms of measuring the code efficiency, we consider its execution time, meanwhile, there may be additional factors to consider, such as memory usage, I/O operations, and the underlying OS environment. Future work may incorporate these factors to perform a more holistic assessment of program efficiency. Lastly, beyond predicting the efficient code, future work may explore its interpretability, providing the reasons why certain codes are more efficient than others at a more fine-grained level (e.g., lines of codes).

Ethic Statement

We believe this work does not have particular concerns about ethics. This is because, it strictly focuses on the technological aspect of comparing code efficiency, which does not engage in the unethical use of LLMs for manipulating software codes, user data, or any other sensitive information.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Singh Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio C'esar Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allison Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xianmin Song, Olatunji Ruwase, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Cheng-Yuan Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv*, abs/2404.14219.

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adeola Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang

Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Federico Cassano, John Gouwar, Francesca Lucchetti, Claire Schlesinger, Carolyn Jane Anderson, Michael Greenberg, Abhinav Jangda, and Arjun Guha. 2023. Knowledge transfer from high-resource to low-resource programming languages for code llms. *ArXiv*, abs/2308.09895.

Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. 2023a. Codet: Code generation with generated tests. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, Jared Kaplan, Harrison Edwards, Yura 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, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Babuschkin, Suchir Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374.

Zimin Chen, Sen Fang, and Monperrus Martin. 2023b. Supersonic: Learning to generate source code optimizations in c/c++. *ArXiv*, abs/2309.14846.

Yihong Dong, Ji Ding, Xue Jiang, Zhuo Li, Ge Li, and Zhi Jin. 2023a. Codescore: Evaluating code generation by learning code execution. *ArXiv*, abs/2301.09043.

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2023b. Self-collaboration code generation via chatgpt. *ArXiv*, abs/2304.07590.

Spandan Garg, Roshanak Zilouchian Moghaddam, Colin B. Clement, Neel Sundaresan, and Chen Wu. 2022. Deepperf: A deep learning-based approach for improving software performance. *ArXiv*, abs/2206.13619.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. Deepseek-coder: When the large language model meets programming - the rise of code intelligence. *ArXiv*, abs/2401.14196.

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, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nourhan Fahmy, Urvashi Bhattacharyya, W. Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jana Ebert, Tri Dao, Mayank Mishra, Alexander Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean M. Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023. Starcoder: may the source be with you! *ArXiv*, abs/2305.06161.

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 Jaymin Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de, Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode. *Science*, 378:1092 – 1097.

Aman Madaan, Alex Shypula, Uri Alon, Milad Hashemi, Parthasarathy Ranganathan, Yiming Yang, Graham Neubig, and Amir Yazdanbakhsh. 2023. Learning performance-improving code edits. *ArXiv*, abs/2302.07867.

Mayank Mishra, Matt Stallone, Gaoyuan Zhang, Yikang Shen, Aditya Prasad, Adriana Meza Soria, Michele Merler, Parameswaran Selvam, Saptha Surendran, Shivdeep Singh, Manish Sethi, Xuan-Hong Dang, Pengyuan Li, Kun-Lung Wu, Syed Zawad, Andrew

Coleman, Matthew White, Mark Lewis, Raju Pavuluri, Yan Koyfman, Boris Lublinsky, Maximilien de Bayser, Ibrahim Abdelaziz, Kinjal Basu, Mayank Agarwal, Yi Zhou, Chris Johnson, Aanchal Goyal, Hima Patel, Yousaf Shah, Petros Zerfos, Heiko Ludwig, Asim Munawar, Maxwell Crouse, Pavan Kapanipathi, Shweta Salaria, Bob Calio, Sophia Wen, Seetharami R. Seelam, Brian M. Belgodere, Carlos Fonseca, Amith Singhee, Nirmit Desai, David Cox, Ruchir Puri, and Rameswar Panda. 2024. Granite code models: A family of open foundation models for code intelligence.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2023. 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.

Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, Duyu Tang, M. Zhou, Ambrosio Blanco, and Shuai Ma. 2020. Codebleu: a method for automatic evaluation of code synthesis. *ArXiv*, abs/2009.10297.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, I. Evtimov, Joanna Bitton, Manish P Bhatt, Cristian Cantón Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre D'efossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code llama: Open foundation models for code. *ArXiv*, abs/2308.12950.

Baptiste Rozière, Marie-Anne Lachaux, Lowik Chanussot, and Guillaume Lample. 2020. Unsupervised translation of programming languages. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sijie Shen, Xiang Zhu, Yihong Dong, Qizhi Guo, Yankun Zhen, and Ge Li. 2022. Incorporating domain knowledge through task augmentation for front-end javascript code generation. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022, Singapore, Singapore, November 14-18, 2022*, pages 1533–1543. ACM.

Atsushi Shirafuji, Md. Mostafizer Rahman, Md. Faizul Ibne Amin, and Yutaka Watanobe. 2023. Program repair with minimal edits using codet5. In *12th International Conference on Awareness Science and Technology, iCAST 2023, Taichung, Taiwan, November 9-11, 2023*, pages 178–184. IEEE.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. 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, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *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*.

Pengcheng Yin and Graham Neubig. 2018. TRANX: A transition-based neural abstract syntax parser for semantic parsing and code generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 7–12. Association for Computational Linguistics.

Zhaojian Yu, Xin Zhang, Ning Shang, Yangyu Huang, Can Xu, Yishujie Zhao, Wenxiang Hu, and Qiufeng Yin. 2023. Wavecoder: Widespread and versatile enhanced instruction tuning with refined data generation. *ArXiv*, abs/2312.14187.

E. Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. 2023. Self-taught optimizer (stop): Recursively self-improving code generation. *ArXiv*, abs/2310.02304.

Shuyan Zhou, Uri Alon, Sumit Agarwal, and Graham Neubig. 2023. Codebertscore: Evaluating code generation with pretrained models of code. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13921–13937. Association for Computational Linguistics.

Terry Yue Zhuo. 2024. Ice-score: Instructing large language models to evaluate code. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 2232–2242. Association for Computational Linguistics.

Table 8: Dataset statistics. The first two rows represent the code statistics made by humans and the other rows are the ones made by machines. PL denotes the programming language.

	PL	Training	Validation	Test
Human	Python	14936	761	562
Human	C++	27005	1316	2038
DSC	Python	17079	1046	594
DSC	C++	35666	1290	1949
gpt-3.5	Python	10674	1306	724
CodeQwen	Python	14854	227	207
Granite	Python	20783	650	315

Table 9: Average accuracy results for code improvement classification with order perturbation across multiple different runs, where we report the variance in parentheses.

	All			**Easy**		
	Python	C++	All	Python	C++	All
Zero-shot	50.26 (0.7)	47.02 (5.9)	47.75 (2.9)	48.49 (4.2)	49.94 (3.2)	49.42 (0.2)
Few-shot	50.43 (1.2)	50.00 (2.7)	50.10 (2.3)	49.61 (0.4)	48.32 (0.0)	48.79 (0.1)
Zero-shot CoT	49.70 (0.9)	50.53 (2.7)	50.34 (2.2)	48.94 (0.0)	50.29 (6.6)	49.80 (2.7)
Few-shot CoT	49.27 (3.2)	49.79 (2.3)	49.67 (2.5)	50.62 (2.5)	48.03 (0.0)	48.97 (0.3)
Ours	**71.67** (1.3)	**62.09** (0.0)	**64.25** (0.1)	**77.60** (0.0)	**70.52** (0.3)	**73.09** (0.1)

A Additional Experimental Setups

Dataset Details We report the dataset statistics in Table 8. Note that, in order to obtain the stable code execution result to decide which code is more efficient than the other, we run every refined code with its original code three times and then select the one whose results are consistent across those three runs. In addition, the code execution is performed following the existing setup (Madaan et al., 2023).

Fine-tuning Details We provide details on fine-tuning the efficiency judgment model: we fine-tune the Code LLM (namely, DeepSeek-Coder-Instruct-1.3B) over 10 epochs with a batch size of 16 and a learning rate of 2e-5, and we select the best epoch based on performance on the validation set.

Prompts In Table 11, we provide the prompts used to elicit the Code LLM to refine the code and to predict the code efficiency (in classification and regression settings). For the efficiency prediction problem, we randomly shuffle the sequence of the original and its refined codes.

B Additional Experimental Results

Here, we provide additional experimental results.

Analysis on Bias for Code Sequence In our code efficiency judgment task, we put a sequence of two codes in the input of Code LLMs, and the Code LLMs may have a bias in this sequence (e.g., predicting the code at the last more often). To see whether they have such a bias, we conduct an additional experiment, flipping the order of the code pairs in the input. We report the results in Table 9, and, from this, we observe that there are no such the notable bias in the sequence of codes.

Table 10: The performance difference between the cases of two classes (Degrade and Improve) and three classes (Degrade, Non-Improve, and Improve), with the regression model.

Training Datasets	Two Classes	Three Classes
Python	76.38	65.40
C++	66.69	70.00
Python & C++	68.87	68.97

Qualitative Analysis We provide some example codes in Python and C++ in Figures 4 and 5. From these two examples, we observe that, despite the difference in grammar across different programming languages, code pairs from them can share the same underlying algorithms. This result supports our finding on generalization ability that our model trained on one programming language can be generalizable to other languages (See Table 4).

Two Classes Analysis For an auxiliary analysis, we measure the performance of the proposed regression model on the dataset with three different prediction classes of Degrade, Non-Improve, and Improve, unlike the one with two classes of Degrade and Improve in our main experiment. As shown in 10, we observe that the overall performance between two classes (Degrade and Improve) and three classes (Degrade, Non-Improve, and Improve) cases is similar. Furthermore, for Python, the performance decreases with the case of three classes, meanwhile, for C++, the performance increases that may be due to the fact that most code pairs for C++ belong to the Non-Improve class (Figure 1) and our model accurately identifies this.

Table 11: A list of prompts that we used for code refinement and efficiency predictions. It is worth noting that the variable inside the parentheses { } is replaced with its actual code.

Types	**Prompts**
Code Refinement	Update the given code to make it more efficient. {Original code}

Efficiency Classification	Given a selection of code, determine which one is the most efficient in computing.
	A: {Original code or Refined code}
	B: {Refined code or Original code}

Efficiency Regression	Given two sets of code, assess how much Code B has improved compared to Code A.
	A. {Original code}
	B. {Refined code}

```
# Python Example
N = int(eval(input()))
print(((N*(N-1))//2))
```

```
// C++ Example
#include<iostream>
using namespace std;
int main() {
        long long int n;
        cin>>n;
        cout<<n*(n-1)/2<< endl;
}
```

Figure 4: Generated Python and C++ samples for the question "For an integer N, we will choose a permutation $\{P_1, P_2, ..., P_N\}$ of $\{1, 2, ..., N\}$. Then, for each $i = 1, 2, ..., N$, let M_i be the remainder when i is divided by P_i. Find the maximum possible value of $M_1 + M_2 + \cdots + M_N$. Constraints N is an integer satisfying $1 \leq N \leq 10^9$".

```
# Python Example
from math import floor,ceil

X = int(eval(input()))
cash = 100
count = 0
while cash < X:
    cash=floor(cash*1.01)
    count += 1

print(count)
```

```
// C++ Example
#include <bits/stdc++.h>
using namespace std;
int main() {
    long long X;
    cin >> X;

    int year=0;
    long long s=100;

    while(s<X){
        s=s*1.01;
        year++;
    }
    cout << year << endl;
}
```

Figure 5: Generated Python and C++ codes for the question "Takahashi has a deposit of 100 yen (the currency of Japan) in AtCoder Bank. The bank pays an annual interest rate of 1% compounded annually. (A fraction of less than one yen is discarded.) Assuming that nothing other than the interest affects Takahashi's balance, in how many years does the balance reach X yen or above for the first time?".

User

We compared the efficiency in computing of codes A and B. As a result, code A was more efficient than B. The improvement fraction is 1.7. Explain why this might be the case.

```
# Code A
n = input()
total = 0
for i in range(2, n+1, 2):
    total += i
```

```
# Code B
n = input()
total = 0
for i in range(1, n+1):
    if i % 2 == 0:
        total += i
```

Response

The efficiency comparison between Code A and Code B, and why Code A is 1.7 times more efficient than Code B, can be explained based on how each code is structured to perform its task. The task here is to compute the sum of all even numbers up to a given nnn. Here's a breakdown of the differences:

```
# Code A
n = input()
total = 0
for i in range(2, n+1, 2):
    total += i
```

- Range Function: This code utilizes the `range()` function with three arguments: the start, stop, and step. It starts at 2 and increments by 2 each iteration, effectively looping only over even numbers.
- Loop Count: The number of iterations in this loop is directly halved compared to Code B, because it increments by 2. For a maximum number nnn, this loop executes approximately n/2n/2n/2 times.
- Conditional Checks: There are no conditional checks inside the loop. Each iteration unconditionally adds the loop variable iii (which is guaranteed to be even) to `total`.

```
# Code B
n = input()
total = 0
for i in range(1, n+1):
    if i % 2 == 0:
        total += i
```

- Range Function: Here, the `range()` function is used with just the start and stop, defaulting to a step of 1. It iterates over every integer from 1 to nnn.
- Loop Count: The loop iterates nnn times, which is twice as many as Code A for the same nnn.
- Conditional Checks: Inside the loop, there is a conditional check `(if i % 2 == 0)` to determine whether each number is even before adding it to total. This conditional check adds computational overhead for each iteration.

Why Code A is More Efficient

1. Fewer Iterations: Code A runs fewer iterations—only about half as many as Code B. This reduces the number of operations (additions and assignments) executed.
2. No Conditional Logic: Code A does not use any conditional logic inside its loop, which saves the time required for evaluating the condition for each iteration in Code B.
3. Optimized Looping: Directly iterating over only the numbers of interest (even numbers) avoids unnecessary iterations over odd numbers, which Code B does but doesn't process.

The improvement fraction of 1.7 suggests that Code A is about 70% more efficient than Code B. This could be measured in terms of execution time or the number of operations performed, and it results largely from the reduced number of iterations and the elimination of conditional checks within each iteration.

Figure 6: A toy example of an explainability of our approach.

User

We compared the efficiency in computing of codes A and B. As a result, code B was more efficient than A. The improvement fraction is 3.02. Explain why this might be the case.

```python
# Code A
from math import factorial as f
n = int(eval(input()))
print((f(n) % (10 ** 9 + 7)))
```

```python
# Code B
def factorial(n, mod=10**9 + 7):
    f = [0]*(n+1)\n f[0] = 1
    for i in range(1, n+1):
        f[i] = (f[i-1]*i) % mod
    return f[n]
n = int(input())
print(factorial(n))
```

Response

The improvement in efficiency from code A to code B by a factor of 3.02 can be attributed to several reasons, primarily concerning how each code handles the calculation of the factorial and the management of large integers.

1. Handling of Large Numbers
Code A uses Python's built-in `factorial` function from the `math` library. This function computes the factorial of n as a single large integer, which can become extremely large as n increases. Only after calculating this massive number does it apply the modulo operation. This approach is memory and computation-intensive because it deals with very large numbers up to the final computation.
Code B takes a different approach by iteratively calculating the factorial and immediately applying the modulo operation at each step. This significantly reduces the size of the integers the program must handle at any one time. By ensuring that no intermediate value exceeds `10**9 + 7`, Code B avoids the overhead associated with manipulating very large numbers.

2. Efficiency in Modulo Operation
Immediate Modulo Application: In Code B, by applying the modulo operation in each step (`f[i] = (f[i-1]*i) % mod`), the size of the numbers involved in the calculations remains bounded by mod. This is a more efficient use of space and typically faster to compute, as each operation is simpler and avoids the potential overhead of managing and operating on very large integers.

3. Space Complexity
Code B also explicitly manages an array f which stores all the factorial values modulo `mod` up to `n`. This allows not only for efficient calculation of `n! % mod` but potentially could be adapted to handle multiple queries more efficiently by reusing precomputed factorials if needed.

4. Use of Python's Built-in vs Custom Function
Optimization: Python's built-in functions are highly optimized; however, when it comes to specific tasks like calculating a factorial modulo a number, a custom-tailored approach like in Code B can outperform by optimizing for the particular constraints (like the modulo operation being used after each multiplication).

5. Potential Overhead in Code A
The `eval` function used in Code A is generally slower and less safe than a direct `input()` function as used in Code B. While the impact on performance here is minimal compared to the factorial calculation itself, it's another aspect where Code B is more optimized. In summary, Code B is more efficient due to its continuous management of number size through the modulo operation, reducing the computational and memory overhead by avoiding the manipulation of extremely large numbers. Additionally, the direct approach without the overhead of unnecessary functions (like `eval`) contributes to its higher efficiency.

Figure 7: A toy example of an explainability of our approach.