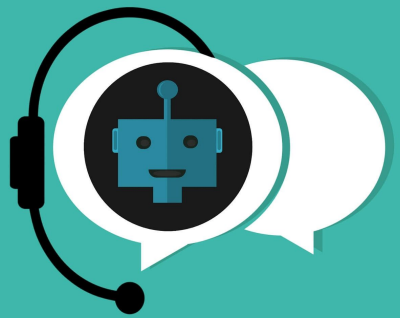
An Evaluation of ChatGPT's ability on generating code solutions for LeetCode Questions



Group 5
- Shiyu Xiu
- Yuangan Zou
- Qian Tang

Motivation

- Explore the possibility of applying ChatGPT as a code-writing assistant during development
- Evaluate the quality of code generated by ChatGPT using Python, Java, C
- Code generation for Leetcode questions provide insights for ChatGPT’s ability in solving real-world questions

Research Questions

- *How does the ChatGPT perform in terms of generating correct code solutions to leetcode questions with different languages?*
  - *Out of 100 questions, How many of the generated solutions are correct and accepted?*

- *How does the ChatGPT perform in terms of running time and memory consumed when solving leetcode questions with different languages?*

Research Questions Cont

- *How does the ChatGPT perform in terms of **Cyclomatic and Cognitive complexity** of the code when solving leetcode questions with different languages?*
    - **Cognitive Complexity:** a measure of how difficult a unit of code is to intuitively understand.
    - **Cyclomatic complexity:** a count of the linearly independent paths through source code.

- *Under what circumstances does ChatGPT **produce an incorrect answer** to Leetcode questions more often?*
    - *hard questions, new questions, complex questions*

Method

A case study with quantitative analysis:

- % of correct answers
- Comparison of running time and memory consumed among languages
- Analysis of readability with a human-assigned score (1 ~5)

Quantitative analysis applied since:

- involve numerical data (# of (in)correct answers)
- statistical methods required

Data collection

Data mining technique: using python selenium library to automatically crawl the leetcode website to obtain the list of problems and its associated metadata such as difficulty, problem description.

Data Analysis

Quantitative analysis:

- Quantitative analysis on the correctness of code solutions
- Evaluate Cyclomatic and Cognitive Complexity with metrics
- Conduct a within-group survey-like study for the readability with human-assigned score (1 ~ 5)

Expected Results

- Table 1 showing : problem info
  - % passed test cases
  - Time & Space Complexity
  - reasons for failure
- Table 2 showing:
  - Readablity scores
  - Cyclomatic and Cognitive Complexity scores

Initial results

Python:

Problem nu	Problem tit	Problem Co	Acceptance	Difficulty	Answers by ChatGPT	Succeeded	Runtime	runtime_be	memory	memory_b	error_type	error_message	Total_testc	Test_cases_passed
1	Two Sum	Given an ar	49.6%	Easy	class Solution: def twoSum	TRUE	61	79.46	15.1	37.43	None	None	None	None
2	Add Two Nu	You are give	40.3%	Medium	# Definition for singly-linked	TRUE	64	85.11	14	32.77	None	None	None	None
3	Longest Sul	Given a stri	33.8%	Medium	class Solution: def lengthO	TRUE	61	76.93	13.9	88.64	None	None	None	None

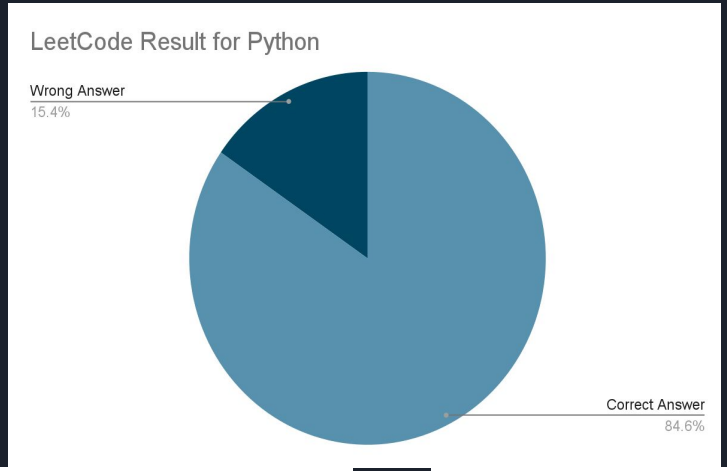
C:

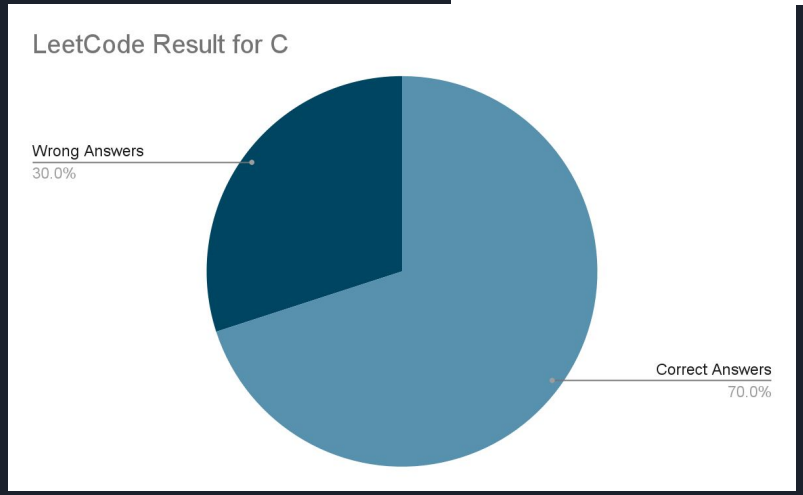
Problem nu	Problem tit	Problem Co	Acceptance	Difficulty	Answers by	Succeeded	Runtime	runtime_be	memory	memory_b	error_type	error_mess	Total_testc	Test_cases_passed
1	Two Sum	Given an ar	49.7%	Easy	/** * Note:	TRUE	132	65.42	6.3	91.3	None	None	None	None
2	Add Two Nu	You are give	40.3%	Medium	/** * Defini	TRUE	19	24.99	7.9	32.88	None	None	None	None
3	Longest Sul	Given a stri	33.8%	Medium	int lengthO	TRUE	3	87.48	5.9	46.69	None	None	None	None

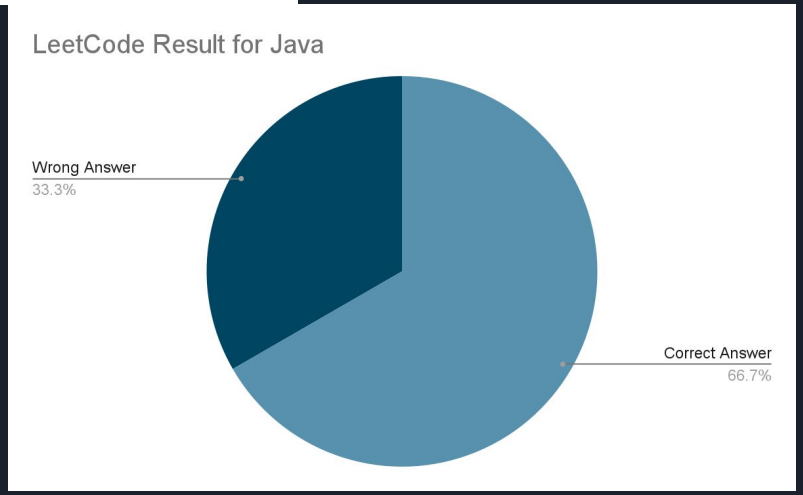
Java:

Problem nu	Problem tit	Problem Co	Acceptance	Difficulty	Answers by ChatGPT	Succeeded	Runtime	runtime_be	memory	memory_b	error_type	error_mess	Total_testc	Test_cases_passed
1	Two Sum	Given an ar	49.7%	Easy	class Solution { public int[] twoSum(int[] nums,	TRUE	1	99.31	43	22.77	None	None	None	None
2	Add Two Nu	You are give	40.3%	Medium	Solution:/** * Definition for singly-linked list. * p	TRUE					compile er	compile er	None	None
3	Longest Sul	Given a stri	33.8%	Medium	class Solution { public int lengthOfLongestSubst	TRUE	2	99.94	42.7	48.66	None	None	None	None

Analysis of Initial Results

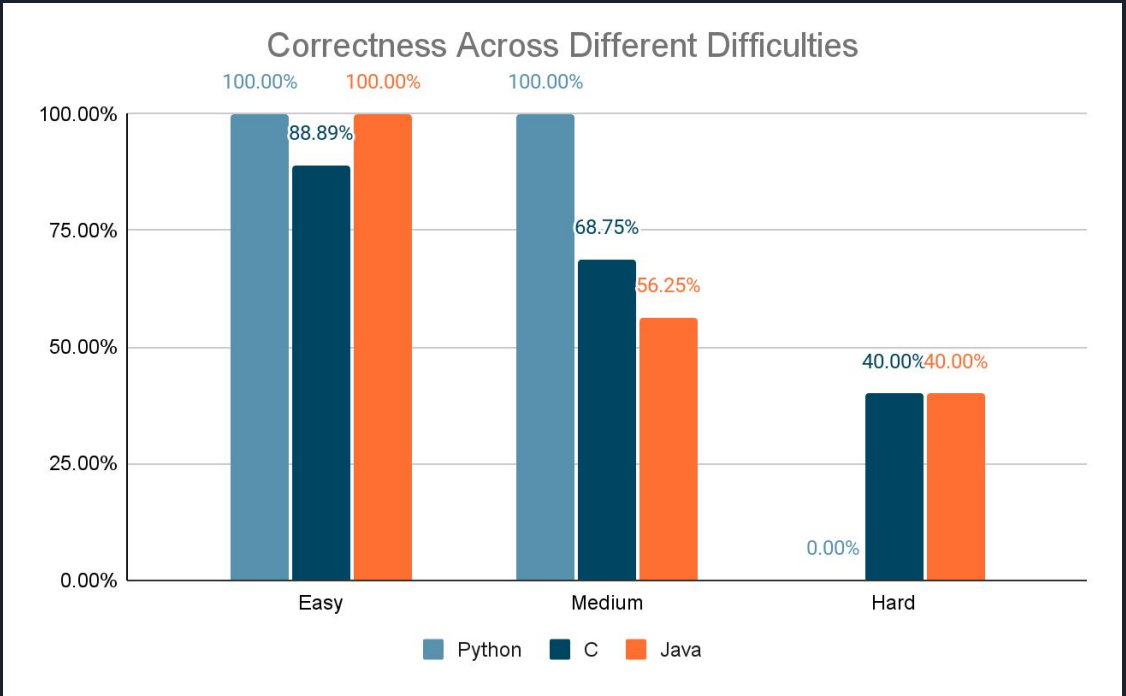






Analysis of Initial Results cont.

% correct answers across different difficulties.



Implications & limitation

- Better understand the ability of cutting edge AI in code generation.
- Help software developers decide whether to use AI tools for writing code.
- Ways of improvement for ChatGPT as a code writer
- Compatibility
- Design of functions

Thank you！