The Behavior of Large Language Models When Prompted to Generate Code Explanations

Priti Oli, Rabin Banjade, Jeevan Chapagain, Vasile Rus
{poli,rbnjade1,jchpgain,vrus}@memphis.edu
University of Memphis, Memphis ,TN, USA

THE UNIVERSITY OF MEMPHIS.

Introduction

- **Code Explanation Generation**
  - Important task across various domains: Software Engineering, Computer Science Education..
  - LLMs for code explanation Generation: Code summarization, Comment generation.
- **Code Explanation in Computer Science Education:**
  - Effective in teaching programming to Novices[1].
  - Self-Explanation of code induces learning gain.
  - Authoring questions, examples and assessment
- LLMS for code explanation generation in educational context[2]
- **Discrepancy in Code Explanations** across different settings:
  - Prompt, Temperature, Programming Language, Explanation Types, LLM settings

Methodology

Variations in Code Explanation Generated by LLM:

- **Intro to Programming Examples**:
  - Programming Languages: Java, Python, C++
  - Code Difficulty: Beginner, Intermediate, Advanced
- **LLMs**:
  - ChatGPT-3.5(ChatGPT-3.5-turbo-0613)
  - ChatGPT-4.0(ChatGPT-4-0613)
  - LLAMA2 (LLMa2-chat)
- **Temperature**: 0, 0.5, 1, 1.5, 2.0
- **Prompt Variations**: 12 different types of prompts (Table 1)
- Total of 3510 explanations

Evaluation

- **Quantitative Evaluation**:
  - Surface level Properties of Explanations:
  - Sentence length, word length, readability (Flesch-Kincaid grade), lexical density, and vocabulary.
- **Qualitative Evaluation**
  - **Accuracy:** Correctness[3]
  - **Completeness:** Information coverage[3]
  - **Conciseness:** Brevity[3]
    Specificity: Alignment with specific code examples
  - Annotated by 2 Graduate Students in binary scale (0/1).

Sym	Prompt	Comment
P1	Can you explain this code?	Simple prompts and Variations
P2	Can you self-explain this code?	
P3	Can you explain this code to a learner?	
P4	Can you explain this code to someone learning to program?	
P5	Can you summarize this code?	Prompts for Summary
P6	Can you summarize this code for a learner?	
P7	Can you explain this code at statement level?	line-by-line explanation
P8	Can you explain this code at block level?	logical/functional explanation
P9	Can you explain this code without breaking down individual statements?	
C1	*<Context1>*.Given this Java code, explain the code to your students in order to help them understand what the code does and learn the covered programming concepts.	Contextual-ized prompt
C2	*<Context2>*.Given this Java code, read the code carefully and explain what it does to potential students who learn programming.	
D1	Explain the following code line by line as a bulleted list:	Prompts used in prior work
D2	Give a detailed explanation of the purpose of the following code.	
D3	Summarize and explain the goal of the above code.	

Table 1: The input prompts used from simple to contextualized.

Experiment and Results

Quantitative Analysis

- Variation with Input Parameter (prompt, code example etc.)
  - Vocabulary,
  - Token length
  - Sentence length
- Prompt C1 and D2 generate significantly longer explanations.
- Readability consistent for Python and Java unlike C++
- Lexical density remains consistent (0.47 on average).

Qualitative Analysis

- Accuracy: 93%
- Completeness: 82%
- Concision: 58%
- Specificity: 77%
- Variations with input parameter in Table 2.

Factors	Values	Complete	Correct	Concise	Specific
Prompt	P1	0.92	0.92	0.54	0.85
	P2	0.85	1.00	0.62	1.00
	P3	0.86	0.79	0.57	0.64
	P4	1.00	0.92	0.42	0.88
	P5	0.86	0.43	0.93	0.57
	P6	1.00	0.86	0.93	0.79
	P7	1.00	0.86	0.39	0.86
	P8	1.00	0.91	0.36	0.82
	P9	1.00	0.60	0.67	0.47
	C1	0.86	0.86	0.43	0.71
	C2	1.00	0.92	0.62	1.00
	D1	0.86	0.79	0.64	0.75
	D2	0.96	1.00	0.46	1.00
	D3	1.0	0.83	0.62	0.88
Temperature	0	0.98	0.81	0.62	0.84
	0.5	0.96	0.91	0.56	0.81
	1	0.93	0.84	0.60	0.82
	1.5	0.71	0.44	0.41	0.35
Model	gpt-3.5-turbo	0.97	0.81	0.75	0.82
	gpt-4	0.96	0.82	0.56	0.88
	Llama2	0.88	0.82	0.41	0.66
Language	Java	0.95	0.80	0.66	0.80
	Python	0.91	0.78	0.43	0.83
	CPP	0.95	0.87	0.66	0.71
Code Example	AreaOfCircle	0.91	0.89	0.49	0.92
	AvgOfNumbers	0.96	0.80	0.50	1.00
	Point	0.94	0.83	0.59	0.81
	BingoBoard	0.93	0.83	0.83	0.80
	BinarySearch	0.96	0.77	0.67	0.66

Table 2: Qualitative Evaluation Scores Across various factors

Discussion and Conclusion

- Explanations for the same prompt vary across example, possibly influenced by the diverse instances available in training data.
- LLMs exhibit diversity/inconsistency in generated explanations based on input parameters.
- GPT-4 generates better explanation than ChatGPT 3.5 and Llama2.
- LLMs' diversity/inconsistency necessitates well-documented parameters and human effort for refining generated explanations.

References

1. Oli, Priti, et al. "Improving Code Comprehension Through Scaffolded Self-explanations." International Conference on Artificial Intelligence in Education. Cham: Springer Nature Switzerland, 2023.
2. MacNeil, Stephen, et al. "Generating diverse code explanations using the gpt-3 large language model." Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 2. 2022.
3. Sridhara, Giriprasad, et al. "Towards automatically generating summary comments for java methods." *Proceedings of the 25th IEEE/ACM international conference on Automated software engineering*. 2010.