Stack Overflow vs. ChatGPT: A Comparative Analysis of Solutions

Gia Phu Tran, Scott Barnett, Ben Philip, Anj Simmons, Anupam Chaudhuri

April 2023

Abstract

This paper presents a comprehensive study on the performance of OpenAI's GPT-3.5-turbo as a programming assistant using Stack Overflow data. We developed an automated methodology for collecting and comparing the AI model's responses with accepted Stack Overflow answers. Our dataset comprised questions posted after 2022 to ensure that the questions were out-of-sample for the AI model, which was trained on data up to 2021. Using the OpenAI Evals framework, we conducted systematic evaluations of the AI model's responses against the accepted human answers. We categorized the results into subsets, supersets, exact matches, disagreements, and non-conflicting differences. The majority of the AI's responses were subsets of the human answers, demonstrating a high degree of factual accuracy. In cases of disagreement or difference, we conducted human evaluations, revealing a preference for GPT's responses for their detail and clarity. Despite these promising results, the exact alignment of AI and human responses was rare, highlighting an area for future improvement. Moreover, our study was limited by focusing on previously answered questions and a single version of the AI model. Future research should explore the performance of AI models on unanswered questions and different versions of the model, and further validate these findings with larger-scale human evaluations. This study contributes valuable insights into the performance of AI models in real-world programming tasks, with implications for the future development of AI-based programming assistants.

1 Introduction

In today's quickly changing technological scene, developers frequently require support when confronted with hard programming issues. A number of resources have been created to meet this requirement, with Stack Overflow and Generative Pre-trained Transformers (GPT) emerging as important platforms for offering technical help and promoting knowledge transfer. In this study, we will evaluate these two platforms and analyse their usefulness in assisting developers in their hunt for programming answers.

OpenAI's Generative Pre-trained Transformers (GPT) are a series of cutting-edge machine learning models, with GPT-4 being the most recent edition. These models have shown outstanding natural language processing and generating abilities, allowing them to participate in human-like conversations, answer queries, and give instruction in a variety of fields, including programming.

Stack Overflow, on the other hand, is a well-known online question-and-answer forum geared towards programmers. It was founded in 2008 and has since become an indispensable resource for developers worldwide, providing a massive collection of questions, answers, and conversations on a wide range of programming issues. The platform's success may be credited to its active user community, who share their experience and knowledge to assist one another.

Developers looking for answers frequently resort to these resources to solve programming challenges, as well as to improve their abilities and awareness of new technologies. The usefulness of these platforms in delivering accurate, dependable, and fast support, on the other hand, has yet to be properly evaluated

and compared .

The desire to better understand the merits and limits of both Stack Overflow and GPT-based language models like ChatGPT in resolving programming-related enquiries drove this study. By comparing these platforms, we hope to throw light on their individual benefits and suggest possible areas for development or collaboration in order to increase the assistance provided to developers in their pursuit of knowledge and technical skill.

2 Background

2.1 Language Models and GPT

Large-scale language models (LLMs), such as GPT, have revolutionised natural language processing (NLP) and understanding by their ability to generate human-like text based on the context provided. GPT, developed by OpenAI, is a series of transformer-based models that have been pre-trained on vast amounts of text data, enabling them to generate coherent and contextually relevant responses across various domains, including programming.

Developers have started leveraging these advanced language models to address their programming challenges and seek guidance on technical issues. LLMs like GPT can be employed as virtual assistants, offering on-demand support for code debugging, documentation, or even suggesting best practices and optimisations.

It is worth to mention that Stack Overflow have a temporary policy to banned ChatGPT [6]. In their opinion, while the answers which ChatGPT produces typically *look like* they *might* be good, it is against their core trust policy when users copy and paste information into answers without validating that the answer provided by GPT is correct prior to posting [10]. Even though they claim: "ChatGPT produces have a high rate of being incorrect". They have performed no research to back this claim, and that is an inspired for this research.

2.2 Related Works

Several studies have explored the performance and potential applications of GPT and other LLMs in different domains. Some research has focused on comparing GPT-generated responses to those provided by human experts, evaluating aspects such as accuracy, coherence, and relevance. These studies have predominantly targeted general knowledge domains or specific industries, such as healthcare [2] [4] [5]

There are technical overviews of ChatGPT perform in Leetcode problem [8] and on Computer Science academic environment [3] [1]. But we want to investigate more variety of problems in the industry.

Need to expand the related works

2.3 Limitations in Existing Works

Despite the growing body of research surrounding LLMs and their potential applications, few studies have specifically investigated their efficacy in addressing coding and programming-related inquiries compared to established platforms like Stack Overflow. Furthermore, most of the research has focused on the performance of LLMs in isolation, rather than in a head-to-head comparison with human-generated responses in the context of programming assistance. [8]

Moreover, the potential limitations and challenges of using LLMs, such as ChatGPT, for programming assistance have yet to be comprehensively explored. Understanding these limitations is crucial for developers who rely on such tools and for improving their integration with existing platforms to enhance the overall user experience in programming-related tasks.

In this study, we aim to address these gaps by comparing the performance of ChatGPT and Stack Overflow in the context of programming assistance and by analysing the factors that contribute to their effectiveness. By doing so, we hope to provide valuable insights into the potential applications and limitations of GPT-based language models in the programming domain.

3 Methodology

3.1 Data source

The data collection process starts when we first sample questions from a comprehensive archive of Stack Overflow data provided by Google Cloud BigQuery (in bigquery-public-data.stackoverflow database). Updated quarterly, this dataset reflects the content on Stack Overflow found within the Internet Archive and can be accessed via the Stack Exchange Data Explorer. Table posts_questions will provide all needed information about the question.

3.2 OpenAI Evals

Evals (short for evaluation) is a software framework for creating and running benchmarks for evaluating models open-sourced by OpenAI [9]. The idea: a framework that can perform large requests to LLMs (include GPT-3.5 and GPT-4) based on a dataset of rules and letting it to checking its own work according to our certain metrics by sending an evaluation prompt. In this case, we will only focus on the version of ChatGPT answer was used: GPT-3.5-turbo in 10 April 2023. The evaluation prompt will need to prime the model to answer in such a way that is easily parsable, like yes/no or A/B/C/D/E.

There are different eval templates for different evaluations and can provide many parameters for that eval template [7]. Some of the important parameters include:

- eval_type: How we expect the model to format and reasoning its response to the evaluation prompt. The one we use for this paper is cot_classify (chain-of-thought then classify), as OpenAI claims it typically provides most accurate [7]. The prompt being asked for this type is:

  "First, write out in a step by step manner your reasoning to be sure that your conclusion is correct. Avoid simply stating the correct answer at the outset. Then print only a single choice from choices (without quotes or punctuation) on its own line, corresponding to the correct answer. At the end, repeat just the answer by itself on a new line."

- Prompt: The evaluation prompt, which should take in the model's completion to the original prompt, potentially along with some other information we pass in.

3.3 Execution plan

The following method is proposed as shown in figure 1:

1. **Collect questions:** We chose to focus on questions posted after 2022 as ChatGPT used Stack Overflow's data before 2022 to train their models. Next, we narrowed down to only those questions that have accepted answers, and we limit the size of the dataset to one thousand. This is done using the query presented in listing 1.

2. **Extract GPT's answer:** Using the OpenAI API with GPT-3.5-turbo and OpenAI Evals, we can use the dataset to generate prompts and receive replies. By extract the answer beforehand, we can perform more than 1 evaluation.

3. **Create samples:** Create a sample JSON file which includes questions from Stack Overflow, ChatGPT's answers from step 2 and the human accepted answer.

4. **Evaluation:**

   - Select test and custom template: the evaluation employed is coqa-fact. For the purpose mention in step 2 it is necessary to modify the model-graded file, fact.yml, to accommodate this situation by allowing it to accept the pre-generated ChatGPT response (listing 2).
   - Comparison using OpenAI Evals: using the test mention above to perform comparison between human answer from Stack Overflow and GPT's answer

5. **Process results:** The submitted answer may either be a subset or superset of the expert answer, or it may conflict with it. Determine which

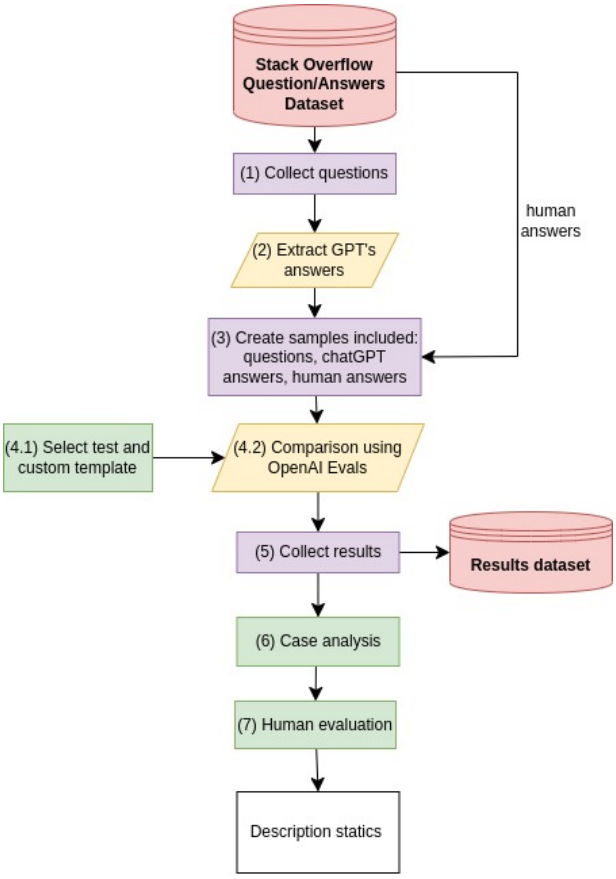

Figure 1: Methodology for data collection and analysis (purple: automate process, red: dataset, green: manual process, yellow: automatic approach using OpenAI API)

case applies. The result will take the form of ABCDE characters to represent that a factual consistency assessment is employed such that, given a response "a" (ChatGPT's answer) and a reference solution "b" (accepted answer from Stack Overflow), the following outcomes are possible:

- Category "A" is assigned when answer "a" is a subset of the expert response "b" and is entirely consistent with it.
- Category "B" is assigned when answer "a" is a superset of the expert response "b" and is entirely consistent with it.
- Category "C" is assigned when answer "a" encompasses all the same information as the expert response "b".
- Category "D" is assigned when there is a disagreement between answer "a" and the expert response "b".
- Category "E" is assigned when answer "a" and the expert response "b" vary, but these differences are not relevant in terms of factual accuracy.

6. **Case Analysis:** We examine how the category being distributed and using the visualisation, statistical tests described in section 4.

7. **Human Evaluation:** We perform human evaluations on the cases where there is a disagreement or difference.

3.4 Procedure for human evaluation

The primary objective of the human evaluation is not to assess the performance of the evaluators themselves, but rather to conduct a side-by-side comparison of the answers provided by ChatGPT and the accepted answers on Stack Overflow on disagreement cases. For this purpose, we employ a choice_scores system to log the frequency with which ChatGPT's answer is judged to be superior to the Stack Overflow accepted answer.

We begin by randomly sampling questions from difference cases and recruiting volunteers who major

in Computer Science to participate in the evaluation process. The volunteers are presented with a pair of responses for each question and are asked to determine "Is the first response better than the second?", without revealing the source of either answer (i.e., whether it is from ChatGPT or Stack Overflow). If the ChatGPT answer is deemed superior, a choice_score of 1 is assigned to that question.

Volunteers are instruct to asset not only the answer that actually work but also many other characteristic of an answer if they unsure which answer is actually works. Some characteristic of it included: the completeness of an answer, clearly explain, details, understandable.

Additionally, we encourage the evaluators to provide written feedback on their choice, explaining why they believe one answer is better than the other. This qualitative input will help us gain a deeper understanding of the factors that contribute to the perceived quality and effectiveness of the answers provided by both ChatGPT and Stack Overflow.

4 Results

4.1 OpenAI Evals evaluation results

The summary for the distribution the evaluations can be seen in this figure 2

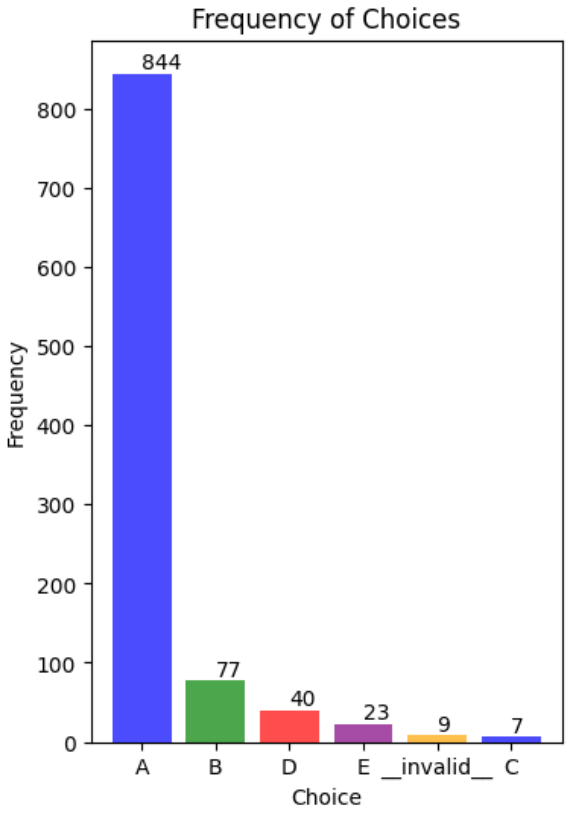

Figure 2: Coqa_fact distribution

In total of 1000 samples, there are 9 cases that not returning valid result for the evaluation here are a table summary for:

Choice	Frequency
Choice = A	844
Choice = B	77
Choice = C	7
Choice = D	40
Choice = E	23

4.2 Human evaluation results

Out of 41 question for human evaluations, volunteers:

1. Prefer human answer 18 times
2. Prefer ChatGPT answer 23 times

5 Discussion

1. How was our work interesting?

   The most frequency result is A (GPT answer is a subset of human answer) show that ChatGPT answer show great alignment with the accepted answer of human for the given question, sometime even superb.

2. What in our results was interesting? Why?

   The frequency of categories in our results is quite intriguing. Category A, where ChatGPT's answer is a subset of the expert response, occurred most frequently (844 times). This suggests that ChatGPT is often able to provide a portion of the correct answer, and potentially solving the problem, though not always the complete response.

   Interestingly, Category B, where the AI response is a superset of the expert response, occurred 77 times. This implies that, at times, ChatGPT is providing more information than what is found in the expert response. While this could be beneficial in some contexts, it could also be seen as overloading the user with information.

   Category D, where there is a disagreement between the AI and the expert response, occurred only 40 times, suggesting a high degree of factual accuracy from the AI. Similarly, category E occurred 23 times, indicating that while the responses may vary but they do not necessarily compromise on factual accuracy and correctness.

   The least frequent category was C, which implies that exact matches between AI and expert responses were rare (7 times) but in these 7 times. We found that answer in this category was exceptionally correct and in alignment with the human answer.

   For human evaluation show if both solution have the same performance, human would perfer ChatGPT answer more for side benefits such as: easy for to understand, have comments, more details, provide example code for the problem and explain why it was wrong. In contract, human answer was better as giving out external source like documentation.

3. What are the implications of our findings in practice?

   Improving future LLMs models: with the results above, we know that for answer fall into category C show exact matches between AI and expert responses, questions and answers fall into this category can be use to improve further AI response in an more automatic approach. In fact we suspect using this approach is how OpenAI engineering team can develop their GPT-4 model so quickly by running their question and answer through a set of many evals and improve the model with the response.

5.1 Limitations

Limitations in our study

The proposed methodology allows us to implement more than one evaluation, so the more evaluation we can do in the future, the more we can understand about the nature of GPT

Our research focuses solely on questions that have already been answered, which may not fully capture the variation in question difficulty. It is possible that there are questions that GPT can answer but humans cannot.

We did not evaluate the difficulty level of the questions or identify the types of questions where GPT excels.

Only focus on GPT-3.5-turbo now

Sometimes ChatGPT reveals that it is an AI in its answer, e.g. uses the phrase "As an AI language model", thus the human evaluation may not be completely blind.

Volunteers did not have time to test every answer thoroughly. Thus, may have sometimes selected an answer that looked correct rather than an answer which actually works.

6 Conclusion

Our study offers insights into the performance of GPT-3.5-turbo in the domain of programming assistance, comparing it to Stack Overflow solutions. A significant number of automated evaluations showed ChatGPT responses were consistent subsets of expert solutions, and in some cases, supersets, providing additional information. However, exact matches were rare, indicating room for better alignment with expert knowledge.

Human evaluations for diffirence cases often favored ChatGPT's answers for their clarity, detail, and explanatory value. The preference for ChatGPT in complex cases indicates its potential in providing effective programming assistance. Nevertheless, further improvements are needed to prevent information overload and to increase the frequency of exact matches with expert answers.

Despite promising results, our study had limitations. We focused on previously answered questions and did not evaluate question difficulty or types where GPT excels. We also only focused on GPT-3.5-turbo. Furthermore, we need to expand our work on the human evaluation.

In conclusion, our study underscores the potential of LLMs like ChatGPT in programming assistance while highlighting areas for improvement and future research. The study serves as a foundation for further exploration in this intersection of natural language processing and programming.

7 Apendix

7.1 SQL lisitng

```
SELECT
  q.id,
  q.title,
  q.body,
  q.accepted_answer_id,
  q.view_count,
  a.body AS stackoverflow_answer
FROM
  bigquery-public-data.stackoverflow.posts_q
↪  uestions
↪  q
LEFT JOIN
  `bigquery-public-data.stackoverflow.posts_
↪  answers`
↪  a
ON
  q.accepted_answer_id = a.id
WHERE
  q.answer_count > 0
  AND q.accepted_answer_id > 0
  AND EXTRACT(YEAR FROM q.creation_date) >=
↪  2022
LIMIT 1000
OFFSET 0;
```

Listing 1: SQL Query

7.2 Model-graded template

```
fact:
  prompt: |-
    You are comparing a submitted answer to an expert
↪   answer on a given question. Here is the data:
    [BEGIN DATA]
    ************
    [Question]: {question}
    ************
    [Expert]: {ideal}
    ************
    [Submission]: {input}
    ************
    [END DATA]
    Compare the factual content of the submitted answer
↪   with the expert answer. Ignore any differences in
↪   style, grammar, or punctuation.
    The submitted answer may either be a subset or
↪   superset of the expert answer, or it may conflict
↪   with it. Determine which case applies. Answer the
↪   question by selecting one of the following options:
    (A) The submitted answer is a subset of the expert
↪   answer and is fully consistent with it.
    (B) The submitted answer is a superset of the expert
↪   answer and is fully consistent with it.
    (C) The submitted answer contains all the same
↪   details as the expert answer.
    (D) There is a disagreement between the submitted
↪   answer and the expert answer.
    (E) The answers differ, but these differences don't
↪   matter from the perspective of factuality.
  choice_strings: ABCDE
  input_outputs:
    question: question_completion
```

Listing 2: Modify Modelgraded: fact.yml

References

[1] Sebastian Bordt and Ulrike von Luxburg. Chatgpt participates in a computer science exam, 2023.

[2] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023.

[3] Ishika Joshi, Ritvik Budhiraja, Harshal Dev, Jahnvi Kadia, M. Osama Ataullah, Sayan Mitra, Dhruv Kumar, and Harshal D. Akolekar. Chatgpt – a blessing or a curse for undergraduate computer science students and instructors?, 2023.

[4] Felipe C. Kitamura. Chatgpt is shaping the future of medical writing but still requires human judgment. *Radiology*, 307(2):e230171, 2023. PMID: 36728749.

[5] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), sep 2022.

[6] Stack Overflow Meta. Temporary policy: Chatgpt is banned, 2023. Accessed: 2023-05-08.

[7] OpenAI. Evaluation templates, 2023. Accessed: 2023-04-30.

[8] OpenAI. Gpt-4 technical report, 2023.

[9] OpenAI. Openai/evals. GitHub repository, 2023.

[10] Stack Overflow. Gpt policy, 2023. Accessed: 2023-05-08.