Are We Ready to Embrace Generative AI for Software Q&A?

Bowen Xu*†, Thanh-Dat Nguyen‡, Thanh Le-Cong‡, Thong Hoang§, Jiakun Liu†,
Kisub Kim†, Chen Gong¶, Changan Niu‖, Chenyu Wang†, Bach Le‡, David Lo†

*North Carolina State University, USA
bxu22@ncsu.edu

†Singapore Management University, Singapore
{bowenxu.2017, jkliu, kisubkim, chenyuwang, davidlo}@smu.edu.sg

‡University of Melbourne, Austrialia
{thanhdatn, congthanh.le}@student.unimelb.edu.au, bach.le@unimelb.edu.au

§CSIRO's Data61, Australia   ¶University of Virginia, USA   ‖Nanjing University, China
james.hoang@data61.csiro.au   fzv6en@virginia.edu   niu.ca@outlook.com

Abstract—Stack Overflow, the world's largest software Q&A (SQA) website, is facing a significant traffic drop due to the emergence of generative AI techniques. ChatGPT is banned by Stack Overflow after only 6 days from its release. The main reason provided by the official Stack Overflow is that the answers generated by ChatGPT are of low quality. To verify this, we conduct a comparative evaluation of human-written and ChatGPT-generated answers. Our methodology employs both automatic comparison and a manual study. Our results suggest that human-written and ChatGPT-generated answers are semantically similar, however, human-written answers outperform ChatGPT-generated ones consistently across multiple aspects, specifically by 10% on the overall score. We release the data, analysis scripts, and detailed results at https://github.com/maxxbw54/GAI4SQA.

I. INTRODUCTION

On November 30, 2022, OpenAI, a world-class AI company, launched an artificial intelligence chatbot named ChatGPT [1]. Since then, it has rapidly become a widely used tool because of its impressive ability to produce responses on various tasks. In just two months after its release, ChatGPT reportedly reached 100 million users, making it the fastest-growing consumer application in history [2]. ChatGPT interacts with users as a chatbot; therefore, the most human-comparable usage of ChatGPT is for question and answering. In the context of software engineering (SE), ChatGPT has already been widely used by programmers to answer technical queries [3]. However, the application was banned by Stack Overflow [4], the largest software Q&A crowdsourcing platform, only after six days from its release. We quote the reason provided by Stack Overflow as follows:

> "Overall, because the average rate of getting correct answers from ChatGPT is too low, the posting of answers created by ChatGPT is substantially harmful to the site and to users who are asking and looking for correct answers."

Despite the above claim from Stack Overflow, there remains no clear empirical evidence on the overall quality of ChatGPT-generated responses as compared to human-written ones on software question answering (SQA). In this booming era of AI-powered chatbots, traffic to OpenAI's ChatGPT has been growing exponentially, while traditional Q&A site such as Stack Overflow has been experiencing a steady decline [5]. Specifically, traffic to Stack Overflow was down by 6% every month in January 2022 on a year-over-year basis and was down 13.9% in March 2022 [6]. This phenomenon, however, is concerning due to the lack of empirical evidence on a comparative study on human-written vs AI-generated responses. The empirical evidence is much needed to ensure a balanced and robust development in the field of SQA. In this work, we investigate the following research questions:

- **RQ1:** *What are the characteristics of ChatGPT-generated and human-written answers?*
- **RQ2:** *From the human user perspective, how good are the ChatGPT-generated answers?*

We answer RQ1 and RQ2 by considering automatic metrics and conducting a manual study, respectively. For RQ1, we find that (1) the average length of answers may not always be consistent with the binary result of which answer is longer than the other, (2) the semantics of human and ChatGPT answers are close to each other, (3) humans and ChatGPT have significantly different opinions on whether the questions should be answered with code snippets or not. For RQ2, we find that (1) human-written answers are still better than ChatGPT-generated answers from 6 aspects (Correctness, Usefulness, Diversity, Readability, Clarity, and Conciseness), (2) ChatGPT-generated answers can fully address only 52% of the questions while human answers can fully address 84%, (3) 27% of ChatGPT-generated answers carry factual errors while only 2% human answers have factual errors. Overall, we accept our hypothesis with both quantitative and qualitative evidence, i.e., ChatGPT-generated answers are semantically similar to human answers; however, they are of lower quality.

Our research result can be useful for future SQA from multiple perspectives, as we describe it in more detail throughout the paper. We find that human users can easily distinguish human-written and ChatGPT-generated answers. It mitigates

the risk of adopting ChatGPT for SQA. Moreover, we find that ChatGPT tends to be conservative in answering different questions differently. We also find that ChatGPT and humans have different opinions on whether a question should be answered with code or not. Nevertheless, the overall quality of ChatGPT-generated answers is still fair and ChatGPT can immediately generate answers which is much faster than waiting for an acceptable answer on Stack Overflow (more than 13 days on average). Therefore, our experimental results suggest that the following directions are worth further research: (1) how to improve the ChatGPT-generated answers, (2) how to design better interplay between humans and ChatGPT for better SQA.

II. DATA PREPARATION

In this section, we describe how we collect technical questions and answers generated by ChatGPT and humans.

A. Technical Question Collection

Following previous works [7–9], we define some criteria to initialize our search space for technical questions on Stack Overflow: (1) As ChatGPT was trained based on the accumulated dataset from September 2021 onwards, we collect questions created in 2022 to mitigate data leak issue; (2) Each question must have an accepted answer; (3) Questions are tagged with a specific programming language. In this paper, we consider the questions tagged as "Java" and "Python"; (4) Each question does not duplicate any other questions; (5) Questions have more than 5 upvotes. Intuitively, the questions with higher votes are more likely to be described clearly, (6) We select questions that do not include images as ChatGPT cannot process images. Based on the selection criteria, our dataset contains 442 and 182 questions related to Python and Java, respectively. The average length, in terms of the number of words, of each Python and Java question is 198 and 225, respectively. Correspondingly, each Python question contains 420 tokens, while each Java question includes 576 tokens.

B. Answer Collection

We consider *accepted answers* on Stack Overflow as these answers, written by humans, are often of high quality [10]. We employ the OpenAI API [11] to query the model *gpt-3.5-turbo*. We set the system prompt (i.e., the role of the ChatGPT in our task) to *You are a software question and answer chatbot for programmers*. And for the user prompt (i.e., the structured content of the question), we use a simple prompt *Question title : [Title] [NEWLINE] Question body : [BODY]*.

III. RQ1: AUTOMATIC COMPARISON

To answer RQ1, we employ a set of automatic metrics on the collected question and answer pairs.

A. Metrics

Length. We consider the *length* of an answer as a proxy for measuring readability and conciseness. Thus, we empirically investigate the length of human-written and ChatGPT-generated answers. Specifically, we calculate the length from

two granularity levels: *words* (appreciated by humans) and *tokens* (considered by ChatGPT). For each answer, we calculate the number of words by considering the space as the delimiter. The GPT family of models process text using tokens, which are common sequences of characters found in text. In this work, we calculate the number of tokens by using the OpenAI tokenizer [12].

Similarity. We measure the similarity between human-written and ChatGPT-generated answers based on their embeddings. That is we map two pieces of text to a semantic vector space and then calculate their distance in the vector space to reflect their similarity. The closer they are, the more semantically similar they are. To implement this, we use the latest embedding model developed by OpenAI, i.e., *text-embedding-ada-002*, and calculate the similarity by following the official instructions [13].

Code recommendation. Considering that code snippets are commonly provided in the answers to technical questions. We investigate if there is a significant difference between humans and ChatGPT in determining whether a question needs to be answered with code or not.

B. Experimental Result

TABLE I: Automatic Comparison between Human and ChatGPT-generated answers

PLs	Answer Type	Avg. # of Words	Avg. # of Tokens	# if longer than the other	Similarity
Python	Human-written	323	387	207	0.86
	ChatGPT-generated	172	236	235	
Java	Human-written	314	345	93	0.86
	ChatGPT-generated	173	219	89	

Table I presents the results of RQ1. On the one hand, we surprisingly find that human-written answers carry more words and tokens than ChatGPT-generated answers. For both Python and Java questions, the human-written answers are around 1.8 times longer than ChatGPT-generated answers. On the other hand, among all the questions considered in this work, 8% of their human-written answers are shorter than ChatGPT-generated answers. It indicates that **the average length of answers may not consistently align with the binary outcome of determining which answer is longer than the other**. Moreover, based on the similarity calculated based on OpenAI text embedding model, we find that the similarity between human and ChatGPT-generated answers to both Python and Java questions only differ slightly. And **overall, the semantics of human and ChatGPT-generated answers are close to each other**. Moreover, we calculate *Cohen's kappa coefficient* [14] to measure the agreement between humans and ChatGPT on determining whether a question should be answered with code or not. We find that the kappa score is only 0.07 which corresponds to the *slight agreement*. It indicates **humans and ChatGPT have significantly different opinions on whether the questions should be answered with code snippets or not**.

IV. RQ2: MANUAL COMPARISON

To answer RQ2, we conduct a manual analysis to evaluate ChatGPT-generated and human-written answers.

A. Participants

We design a customized questionnaire for each question and then distribute it to the participants. In total, we invite 7 participants. For questions related to different programming languages, we assign them to participants who have at least 2 years of programming experience. We ask participants to skip questions for which one or more of these conditions are satisfied: (1) the participants think they are not knowledgeable enough to answer the questions, (2) the participants have read the given question and its corresponding answers before. Moreover, we prohibit participants to search the original Stack Overflow post during the manual study.

B. Questionnaire Design

Inspired by [15], we randomly sort the answers generated by human and ChatGPT across different questionnaires. Thus, our participants may not receive the answers in the same order. Before asking questions, we first present the title and the body of the target technical question. Then, we present an answer either generated by humans or ChatGPT. For each answer, we ask the following questions:

Q1. *How satisfied are you with the answer? (Required)*
Inspired by prior works [16, 17], we consider 6 aspects to measure the quality of an answer on a 5-point Likert scale (from 1: Very dissatisfied to 5: Very satisfied). The 6 aspects are: Correctness, Usefulness, Diversity, Readability, Clarity, and Conciseness. Besides, we also ask participants to measure the overall quality of the given answer.

Q2. *Please explain your rate on the answer.*
Besides the scores in Q1, we encourage participants to provide further explanations for their ratings.

Q3. *Do you think the answer correctly understands the question? (Required)*
◯ Yes ◯ No ◯ Partially
Having a correct understanding is the first and also the key step to coming up with a correct answer. Therefore, we are interested in investigating the capability of ChatGPT on question understanding.

Q4. *Do you think the answer fully addresses the question? (Required)*
◯ Yes ◯ No ◯ Partially
Certainly, the ultimate goal of SQA is to fully address the technical questions. Thus, we evaluate the capability which requires not only understanding the question but also providing the correct solutions with essential explanation.

Q5. *Is there any factual error in the answer? (Required)*
◯ Yes ◯ No
Hallucination has been complained about by ChatGPT users and the issue has also been admitted by the OpenAI technical report [18].

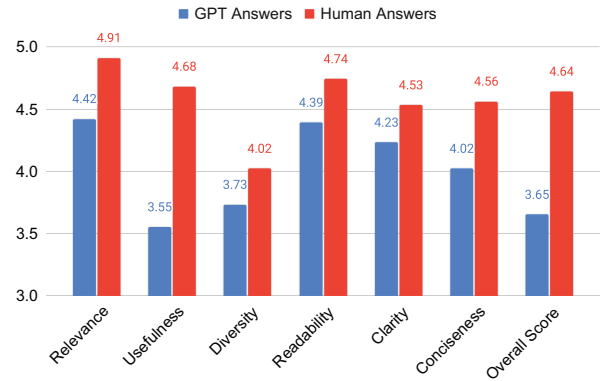**Q6.** *If your answer to the previous question is Yes, please explain.*

Fig. 1: Result of Q1. *How satisfied were you with the answer?*

Q7. *Can you guess which one is generated by AI? (Required)*
◯ Answer #1 ◯ Answer #2 ◯ I cannot recognize.
Whether human users can distinguish ChatGPT-generated content is the key to embracing generative AI techniques. In our case, we are interested in investigating whether humans can easily distinguish human and ChatGPT-generated answers. Thus, after participants answer the above questions for each answer, we ask them to guess which answer is more likely generated by ChatGPT.

C. Result and Analysis

We sampled 20 questions for each of the considered programming languages, i.e., 40 questions are selected in total. And for each question, we assign 2 participants to evaluate the quality of human and ChatGPT generated answers. In total, we create 40 questionnaires. 14 (17%) of the questionnaires are not answered by participants since they think they are not knowledgeable to evaluate the answers to the specific questions. Following are the result and analysis based on the valid responses.

Q1. *How satisfied are you with the answer? (Required)*
Q2. *Please explain your rate on the answer.*

Figure 1 presents the result of Q1. Among the 6 aspects of answer quality measurement, human-written answers are consistently better than ChatGPT-generated answers. However, the gaps for the aspects, Relevance, Readability, and Clarity, are relatively small. 54% of responses show human-written answers are better than ChatGPT-generated answers and only 11% of them show ChatGPT-generated answers are better in terms of the overall score. **On average, human-written answers outperform ChatGPT-generated answers by 10% on the overall score**. From explanations received in Q2, we identify 2 additional reasons why participants scored human-written answers higher than ChatGPT ones: (1) **Generalizability**. ChatGPT fails to generate appropriate answers for questions that do not appear in its training data, which is up to September 2021 [19]. For instance, a question with ID 72166259 demonstrated this limitation: the human-written answer was preferred because *"This answer solves the question with the latest knowledge"*. (2) **Correct-but-useless**

TABLE II: Results of Q3, Q4, and Q5

Question	Human-written Answers			ChatGPT-generated Answers		
	Yes	Partially	No	Yes	Partially	No
Q3. Correctly understands the question?	95%	3%	2%	80%	11%	9%
Q4. Fully addresses the question?	84%	2%	14%	52%	21%	27%
Q5. Any factual error?	2%	NA	98%	27%	NA	73%

content. We find that ChatGPT-generated answers may carry correct-yet-useless information. For example, one participant mentions that *"... although the answer claims it can solve the problem, its content is not even relevant to the problem"*. (3) **Make naive mistakes**. Many works have demonstrated that generative AI has a big potential for software tasks, such as program repair [20–22]. Based on our results, we find that the patch or code change suggestions generated by ChatGPT may carry simple errors that can be easily identified by humans. For example, one participant mentions that *"... The code given seems to just repeat the code given in the question body with some minor modifications. Arrays::stream is not applicable to the stream of char[], so the code given can not be compiled and thus can not solve the question"*.

Q3. *Do you think the answer correctly understands the question? (Required)*

From Table II, we find that the number of human-written answers which fully understand the question is 15% more than the number of ChatGPT-generated answers. However, there are still 80% of ChatGPT-generated answers fully understood the question. It indicates that **ChatGPT has achieved a desirable capability on question understanding, although still not as good as humans.**.

Q4. *Do you think the answer fully addresses the question? (Required)*

From Table II, we find that human-written answers are better than ChatGPT-generated answers in fully addressing the questions by a large margin (32%).

Q5. *Is there any factual error in the answer? (Required)*

Q6. *If your answer to the previous question is Yes, please explain.*

From Table II, we find that human-written answers significantly carry fewer factual errors as compared to ChatGPT-generated answers. For example, there is one factual error identified in the ChatGPT-generated answer that says *"This answer says that com.sun.xml.bind:jaxb-impl is no longer being actively developed or maintained, but the fact is that it is still getting new releases. See https://mvnrepository.com/artifact/com.sun.xml.bind/jaxb-impl."*

Q7. *Can you guess which one is generated by AI? (Required)*

We find that for 86% of questions, participants can correctly distinguish human and ChatGPT-generated answers. Their further explanation provides the reason. For example, *"ChatGPT-generated answers without any emotion"* and *"From my experiences with ChatGPT, it usually provides complete and high-level answers with many redundant details... human-written answers are usually short and directly point out the solution"*.

V. THREATS TO VALIDITY

The programming languages considered in our experiments is a threat to external validity. Similar to some prior works (e.g., [17]), we mitigate this threat by considering two popular programming languages, i.e., Java and Python. Nevertheless, replicating our work for other programming languages is required to broaden our understanding of the capability of ChatGPT. Threats to construct validity are related to the used metrics. In this work, we use multiple automatic metrics to show the characteristic of ChatGPT-generated answers. However, considering automatic metrics alone may not be sufficient, we further mitigate this threat by performing both automatic and manual comparisons between human-written and ChatGPT-generated answers.

VI. RELATED WORK

An et al. [23] conducted a case study on whether developers potentially reused code from Stack Overflow. Their studies suggested that developers may have copied the code of Android apps to answer Stack Overflow questions. Fischer et al. [24] and Zhang et al. [25] show these implications when reusing vulnerable code snippets from Stack Overflow. They highlight the security implications that can arise when developers reuse code from the platform without verifying them against potential vulnerabilities. Our study complements these studies by showing how generative AI, specifically ChatGPT can perform SQA support and how it compares to human-written answers as well as investigating the quality of their answers in various aspects linked to the aforementioned issues including correctness, usefulness, etc.

VII. CONCLUSION

In this work, we compare answers generated by ChatGPT and humans in StackOverflow. Our study uncover the following findings: ChatGPT-generated answers are of lower quality than humans for all aspects considered. Still, they are promising and the gap is smaller for diversity, readability and clarity. Our preliminary study sheds light on the potential future of SQA by summarizing the limitations of ChatGPT.

REFERENCES

[1] "Chatgpt," https://openai.com/blog/chatgpt.

[2] "Chatgpt users," https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

[3] "Chatgpt for software q&a," https://www.infoworld.com/article/3689172/chatgpt-and-software-development.html.

[4] "Stack overflow banned chatgpt," https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned.

[5] "Stack overflow traffic," https://stackoverflow.blog/2018/04/26/stack-overflow-isnt-very-welcoming-its-time-for-that-to-change/.

[6] "Stack overflow traffic decline," https://www.similarweb.com/website/stackoverflow.com/#demographics.

[7] F. Calefato, F. Lanubile, and N. Novielli, "How to ask for technical help? evidence-based guidelines for writing questions on stack overflow," *Information and Software Technology*, vol. 94, pp. 186–207, 2018.

[8] M. Ahasanuzzaman, M. Asaduzzaman, C. K. Roy, and K. A. Schneider, "Mining duplicate questions in stack overflow," in *Proceedings of the 13th International Conference on Mining Software Repositories*, 2016, pp. 402–412.

[9] A. Baltadzhieva and G. Chrupała, "Predicting the quality of questions on stackoverflow," in *Proceedings of the international conference recent advances in natural language processing*, 2015, pp. 32–40.

[10] "Stack overflow accepted answers," https://stackoverflow.com/help/accepted-answer.

[11] "Openai text similarity," https://platform.openai.com/docs/api-reference/chat/create.

[12] "Openai tokenizer," https://platform.openai.com/tokenizer.

[13] "Openai text and code embeddings," https://openai.com/blog/introducing-text-and-code-embeddings.

[14] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[15] B. Xu, L. An, F. Thung, F. Khomh, and D. Lo, "Why reinventing the wheels? an empirical study on library reuse and re-implementation," *Empirical Software Engineering*, vol. 25, pp. 755–789, 2020.

[16] B. Xu, Z. Xing, X. Xia, and D. Lo, "Answerbot: Automated generation of answer summary to developers' technical questions," in *2017 32nd IEEE/ACM international conference on automated software engineering (ASE)*. IEEE, 2017, pp. 706–716.

[17] C. Yang, B. Xu, F. Thung, Y. Shi, T. Zhang, Z. Yang, X. Zhou, J. Shi, J. He, D. Han *et al.*, "Answer summarization for technical queries: Benchmark and new approach," in *37th IEEE/ACM International Conference on Automated Software Engineering*, 2022, pp. 1–13.

[18] "Hallucination issue of chatgpt," https://cdn.openai.com/papers/gpt-4.pdf.

[19] "Training data of chatgpt," https://help.openai.com/en/articles/6783457-what-is-chatgpt.

[20] H. Pearce, B. Tan, B. Ahmad, R. Karri, and B. Dolan-Gavitt, "Examining zero-shot vulnerability repair with large language models," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 2022, pp. 1–18.

[21] C. S. Xia, Y. Wei, and L. Zhang, "Automated program repair in the era of large pre-trained language models," in *Proceedings of the 45th International Conference on Software Engineering (ICSE 2023). Association for Computing Machinery*, 2023.

[22] C. S. Xia and L. Zhang, "Less training, more repairing please: revisiting automated program repair via zero-shot learning," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2022, pp. 959–971.

[23] L. An, O. Mlouki, F. Khomh, and G. Antoniol, "Stack overflow: A code laundering platform?" in *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, 2017, pp. 283–293.

[24] F. Fischer, K. Böttinger, H. Xiao, C. Stransky, Y. Acar, M. Backes, and S. Fahl, "Stack overflow considered harmful? the impact of copy&paste on android application security," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 121–136.

[25] H. Zhang, S. Wang, H. Li, T.-H. Chen, and A. E. Hassan, "A study of c/c++ code weaknesses on stack overflow," *IEEE Transactions on Software Engineering*, vol. 48, no. 7, pp. 2359–2375, 2022.