A Comparison of ChatGPT's Stack Overflow Post Extraction Summaries with the SoSum Dataset Summaries

Emmeline Pearson
Ep623@drexel.edu
Drexel University
Philadelphia, PA USA

Abstract—Building on the SoSum data set, which presented extractive summaries of Stack Overflow posts, this study creates a dataset generated by ChatGPT and compares the results between the two datasets. Stack Overflow is a widely used tool among developers, however it can take time to filter through question answers to find the answer that the developer needs. To address this difficulty, the SoSum data set was created. The SoSum dataset presents a possible solution of using extractive summary sentences, of each answer, in order to expedite the process of searching for the needed answer. The SoSum data set had humans create extractive summaries, which is time intensive, so in comparison, this study looks at using ChatGPT to create those same extractive summaries. Compared to the SoSum dataset, the ChatGPT dataset had very little accuracy (average F1 score of 0.28) and ChatGPT had high levels of confidence in its answers (average confidence level of 8.5 out of 10). Based on these results, improvements to prompt selection, and the ChatGPT model on the whole, need to be made before it will be a valid option in creating extractive summaries of Stack Overflow posts at a similar level to the SoSum dataset.

I. Introduction

Stack Overflow is a widely used website among software developers that allows developers to ask questions, share knowledge and work together to solve problems. Stack overflow has over 21 million users and over 5.5 million visits to the site per day [1]. According to the stack overflow developer survey from 2022, 90% of respondents indicated that they visited stack overflow when they were stuck on a coding problem and 62% of respondents spent more than 30 minutes a day searching for answers or solutions to problems [2]. This time spent filtering and searching through Stack Overflow prompted the creation of the SoSum dataset. The SoSum dataset created extractive summaries with the goal of aiding developers in quickly identifying the answers they were looking for. The goal of the creation of this dataset was to be able to create similarly accurate extractive summaries while reducing the time spent creating each summary by using ChatGPT.

ChatGPT is a large language model created by OpenAI which was trained on a very large data set and aided by reinforcement learning with human feedback. Previous studies have shown ChatGPT's ability to summarize text effectively at the level that satisfies human needs [3]. A large language model's success is often dependent on the data set that it is trained on and the specific topic areas where reinforcement learning was applied to. ChatGPT's previous summarization success was not specific to software development related texts, however, there have been previous research studies looking at ChatGPT's application in the field of software development. In the study of ChatGPT's accuracy for answering software development test questions, it answered 77.5% of questions and of those answered 55.6% were partially or fully correct responses [4]. This study will continue that investigation into ChatGPT's application in software development by looking at how accurate ChatGPT is at summarizing Stack Overflow posts.

II. Research Questions

RQ1: When compared to the SoSum dataset, how accurate is ChatGPT at summarizing the same posts?

The first research question looks into comparing the accuracy of ChatGPT's extractive summaries with those created by humans for the SoSum dataset. This will reveal ChatGPT's usefulness as a tool in summarizing the same Stack Overflow question answers.

RQ2: How does ChatGPT's confidence level relate to the accuracy of the summaries it generates?

The second research question looks into ChatGPT's perceived level of confidence. This will be accomplished by asking ChatGPT to rate it's confidence level on a scale of 1 to 10 for the summary it previously created. Previous studies have shown that ChatGPT's confidence has little bearing on its actual accuracy/success [4]. This question looks into if that still holds true with this model of Chat GPT (text-davinci-003).

III. Methodology

The overall process for creating the ChatGPT data set was as follows: preprocessing the SoSum data set, prompt testing, and lastly generating responses from ChatGPT. The first step was to preprocess the SoSum data set. This involved removing non ascii characters and combining the answer and question datasheets which were provided in the SoSum dataset GitHub [5]. Non ascii characters were not currently supported when using the ChatGPT API. The second step was prompt testing. The goal of prompt testing was to pick a prompt that was most effective in getting accurate extractive summary responses from ChatGPT. To conduct prompt testing, a small subset of 50 samples was chosen randomly from the preprocessed SoSum dataset. Using this subset, 8 different prompts were tested out using the ChatGPT API. Analyzing the accuracy and consistency of results the following prompt was chosen: “Choose one (or more) sentence(s) from this text which answers the question: <question>. Your chosen sentence should capture the overall meaning of the answer text. Answer text: <answer response>”. The third step was to create the ChatGPT dataset by using the selected prompt and asking ChatGPT to create the extractive summaries for all the question responses in the SoSum dataset.

Discussion on Prompts—Eight variations of prompts were tested on the initial testing subset. Choosing a prompt was one of the biggest challenges in the creation of the ChatGPT dataset. Initially prompts tested involved the word “summary” however, any inclusion of this word caused ChatGPT to synthesize the text and return a summary which was not extractive or concise. It would often embellish and add extra words/phrases which were not present in the original text it was asked to summarize. Based on these findings, the word “summary” was left out of the prompt used to generate summaries.

Discussion on Rate Limiting—The ChatGPT API uses rate limiting to control the speed at which it handles requests. One major problem faced at the start of creating the ChatGPT data set was handling rate limiting. The request handling script included sleeps of increasing duration, when rate limiting errors were received, in order to handle these cases. The ChatGPT API sets rate limits based on past usage, so during the first month of testing the limit was set low. On further testing, after a month had passed, the rate limit was increased, and less rate limit errors were seen. Also, one point of note was that testing at odd times (such as at night or during the weekend) seemed to have less rate limiting errors in general.

For the confidence level research question, a smaller sample set was used (50 summaries and confidence level questions). The methodology here followed the same process as the larger ChatGPT dataset, except for the addition of the confidence level question. After ChatGPT had generated its summary, each response was followed with this question: "How confident are you in that answer on a scale of 1 to 10?". ChatGPT then returned a value which was saved alongside the summary, so that accuracy level compared to perceived confidence level could be analyzed.

IV. Results and Discussion

Overall, the ChatGPT dataset had 2,491 generated summaries. Of the 2,491 summaries, 491 of them had zero scores in either or all of the categories (Bleu, Rouge and F1). Discarding the 491 summaries that had statistics of 0, table 1 shows the average results for each of the scores. The nltk python library [6] was used to calculate Bleu score which is a measure of precision in accuracy between a reference text, and a sample text. Bleu score is calculated on a scale of 0 to 1, with a score close to 1 being more accurate than a score closer to 0. For the 2000 summary samples the average bleu scores was 0.236. For rouge score, the rouge-score python library was used. Rouge score is a measure of recall, meaning the portion of words in the reference summary that were also in the sample summary. Rouge score is also on a scale of 0 to 1, with 1 meaning a greater number of words were present in the sample summary from the reference summary. The average rouge score was 0.497. F1 scores combine together the bleu and rouge scores to give an overall metric of the accuracy of a model. F1 scores are on a scale of 0 to 1, where values less than 0.5 are not good, values of 0.5 – 0.8 are okay models, values of 0.8 to 0.9 are good models and scores above 0.9 are very good models. For this data set, the average F1 score was 0.278.

Statistic	Average Result (out of 1.0)
Bleu	0.236
Rouge	0.497
F1	0.278

Table 1: Statistics for 2000 summaries generated by ChatGPT.

Looking more deeply into F1 scores, table 2 shows the distribution of the number of scores at different ranges. There were 203 samples(10.1%) which scored above 0.8 in F1 score whereas there were 814 (40.5%) that scored at above 0.3. Similarly, in figure 1, the distribution of scores shows a large concentration right around F1 scores of zero, however, there are some F1 scores that are pretty good in the higher ranges.

F1 Score Range (out of 1.0)	Number Above (out of 2000)
> 0.8	203
> 0.5	492
> 0.3	814

Table 2: F1 Score ranges and corresponding number of summaries.

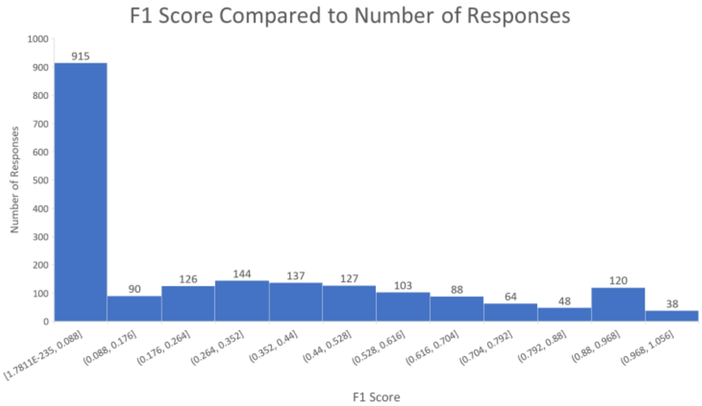

Figure 1: Distribution of F1 scores.

The results show low similarity in the summaries created by ChatGPT compared to those in the SoSum Dataset. The low similarity could be caused by many factors such as: ChatGPT training, choice in prompt, or non-deterministic behavior of ChatGPT. The first factor is that ChatGPT may not have sufficient data or reinforced learning in the areas that the Stack Overflow questions related to. This lack of data or training means that ChatGPT may not have been able to accurately choose a good extractive summary. The second factor is the choice in prompt. This study used one prompt for asking ChatGPT to create extractive summaries, and it is possible that there are other prompts which would have resulted in higher scored extractive summaries. Lastly, the non-deterministic behavior of ChatGPT means that the same input to ChatGPT can result in differing output. This behavior means that it could be possible that on different trials of these same prompts the resulting similarity between the two datasets could be higher (or lower). Although, with a sample size of over two thousand it is unlikely that the average similarity would differ by much due to the non-deterministic behavior of ChatGPT.

Confidence Levels—In the measures of confidence level, 50 samples were taken, and of those 16 had zero scores in one or all of the categories (Bleu, Rouge and F1). For every confidence level question, ChatGPT either rated itself at 8 or 9, with an average level of 8.48. For this smaller sample size, the average bleu score was 0.23, the average rouge score was 0.57 and the average F1 score was 0.27 (see table 3). These scores are similar to the statistics for the larger ChatGPT dataset. Looking more closely at the F1 scores, 4 summaries (8%) had F1 scores over 0.8, 9 summaries (18%) had scores over 0.5 and 18 (36%) had scores over 0.3 (see table 4). In table 5, the average F1 scores are grouped by ChatGPT's perceived confidence levels. Of the 50 overall samples, ChatGPT ranked 27 at a confidence level of 8 and 23 at a confidence level of 9. The confidence level of 8 summaries had an average F1 score of 0.18 and the confidence level of 9 summaries had an average F1 score of 0.21.

Statistic	Average Result (out of 1.0)
Bleu	0.23
Rouge	0.57
F1	0.27

Table 3: Confidence Level Dataset Statistics.

F1 Score Range (out of 1.0)	Number Above (out of 50)
> 0.8	4
> 0.5	9
> 0.3	18

Table 4: Distribution of Confidence Level Summary F1 Scores.

Confidence Score (out of 10)	Average F1 Score (out of 1.0)
8	0.18
9	0.21

Table 5: Confidence Score Compared to Average F1 Score.

Although, the sample size in this case is not to the full extent of the SoSum dataset, these results for confidence level reaffirm what previous studies have shown. ChatGPT's perceived confidence level has little reflection in the accuracy of its answers.

V. Threats to Validity

One limitation of this project was that it only tested the questions and answers in the SoSum data set, which is around 2.3 thousand questions. Using the SoSum data set, as the gold standard to compare to, limited the sample size scope but it provided the human generated summaries to compare to and the initial starting point to assess ChatGPT's accuracy at the task. Secondly, limited prompt testing was conducted. To get a better sense of the accuracy of ChatGPT's summarization abilities on Stack Overflow posts, more prompts could have been used and compared across the whole SoSum dataset. Out of the eight tested prompts, the one selected had the highest accuracy on the initial subset of testing data, however it would have been beneficial to test more initial prompts and to test more chosen prompts on the larger dataset. Thirdly, there were a lot of ChatGPT answers that were invalid and resulted in scores of 0 (bleu, rouge and F1). This could be because ChatGPT didn't answer the prompt, or the API reached the token limit, or the response didn't hold any similarity to the reference response in the SoSum data set. Extensive investigation into each case did not occur, but for future studies this could be an area to investigate. Lastly, the SoSum dataset cut out some code segments by adding "<code>" flags in place of code snippets. It would be interesting to look at how ChatGPT is helped or hurt by the missing information contained in those code snippets.

VI. Conclusion

To address problems Software developers have faced in locating specific Stack Overflow answers quickly, the SoSum dataset was created. The SoSum dataset created extractive summaries of answer post by having humans manually select each chosen sentence. To build upon the SoSum data set, this paper presented the ChatGPT dataset which created extractive summaries of the same questions but instead used ChatGPT to automatically generate the summaries. The research question that was investigated was how these two summary datasets would compare. Based on the results, this created ChatGPT dataset had little similarity to the SoSum dataset. The same level of accuracy to SoSum was not achieved using this prompt and ChatGPT model, the resulting F1 score was 0.278.

In the investigation of confidence ChatGPT's perceived high confidence levels did not correlate to its accuracy in summaries. Many previous studies have shown that large language model confidence levels have little correlation to accuracy of responses, and these results align with that finding.

Further improvements to prompt generation, or the training of ChatGPT, need to be implemented before ChatGPT will be useful as an accurate replacement for humans creating extractive summaries of Stack Overflow posts. In the future these results could be extended by investigating different types of prompts to see if better extractive summaries can be obtained using ChatGPT. Equally, as new models of the ChatGPT API come out this process could be repeated to see if similarity between the SoSum dataset and ChatGPT summaries increases.

VII. References

[1] *Stack exchange*. All Sites - Stack Exchange. (n.d.). https://stackexchange.com/sites?view=list#questions

[2] "Stack Overflow Developer Survey 2022." *Stack Overflow*, survey.stackoverflow.co/2022/#overview.

[3] Zhang, Haopeng, et al. *SummIt: Iterative Text Summarization via ChatGPT.*

[4] Quibria, M. G., et al. "New Information and Communication Technologies and Poverty: Some Evidence from Developing Asia." *Journal of the Asia Pacific Economy*, vol. 7, no. 3, Jan. 2002, pp. 285–309, https://doi.org/10.1080/13547860220000007852. Accessed 3 Nov. 2021.

[5] BonanKou. "SOSum: A Dataset of Extractive Summaries of Stack Overflow Posts and Associated Labeling Tools." *GitHub*, 11 May 2023, github.com/BonanKou/SOSum-A-Dataset-of-Extractive-Summaries-of-Stack-Overflow-Posts-and-labeling-tools. Accessed 10 June 2023.

[6] "Nltk.translate.bleu_score — NLTK 3.6 Documentation." *Www.nltk.org*, www.nltk.org/_modules/nltk/translate/bleu_score.html.

[7] LLC, Google. "Rouge-Score: Pure Python Implementation of ROUGE-1.5.5." *PyPI*, pypi.org/project/rouge-score/. Accessed 10 June 2023.