The Impact of LLM-Based Tools on User Engagement and the Quality of Q&A Discussions on Tech Forums

Sunwoo Baek
University of Illinois Urbana-Champaign

Eisha Peyyeti
University of Illinois Urbana-Champaign

Bridget Agyare
University of Illinois Urbana-Champaign

Himnish Jain
University of Illinois Urbana-Champaign

1 ABSTRACT

This study examines the impact of the introduction of Large Language Models (LLMs) within technical question and answer forums, focusing specifically on Stack Overflow. Given the rise of LLM tools like ChatGPT since 2022, we were motivated to explore whether users are becoming less engaged on these platforms, since LLMs can provide, often, quicker answers. In addition, we wanted to explore how the presence of AI-generated answers in the Stack Overflow community impacts community dynamics. We analyze sets of 200000 and 2000 randomly sampled Stack Overflow responses to posts from 2022-2024. Our analysis employed GPT-based classification, sentiment analysis, and multivariate regression modeling. Our findings indicate that while LLM-generated content is growing, it still compromises a minority of total responses to questions on the platform. We found that comments written in the post-LLM era received significantly fewer upvotes on average, though overall effect sizes were modest. Human written answers continue to receive more upvotes and trust from the community, though statistical differences between human and AI-generated content were not significant. Overall, these findings highlight the growing influence of AI in online discussions, and the importance of maintaining quality in knowledge-sharing platforms. As LLM usage continues to expand, continued monitoring is essential to preserve the authenticity and helpfulness on Stack Overflow, and similar platforms as well.

2 INTRODUCTION

The rapid advancement of Large Language Models (LLMs) has affected many aspects of everyday life. Specifically, it has transformed how users interact within online Q&A forums, particularly in technical domains. Platforms that thrived on user-driven discussions and collaborative problem-solving now face an increasing reliance on AI-generated responses. LLMs like ChatGPT can quickly generate answers, reducing the effort required to participate in discussions. However, these might discourage users' diverse perspectives, decrease users' participation, and change the ways people contribute and exchange knowledge on the forums. In the past, where LLMs weren't used as much as they are currently, users would actively ask questions, discuss solutions, debate their approaches, and refine answers collaboratively. With LLM-generated responses, users might passively consume AI-generated answers instead of engaging in discussion, reducing the depth and diversity of responses. Accordingly, fewer people may feel motivated to contribute, leading to a decline in community-driven learning. If user participation declines, forums risk losing their role as spaces for interactive learning and dynamic discourse, potentially impacting the very ecosystems that AI models rely on for fresh problem-solving data. Motivated by this, this study examines the effects of LLM-based tools on user engagement, perceived helpfulness, and the overall quality of Q&A discussions. This project aims to provide insights into the evolving role of AI in online knowledge-sharing communities. The advent of generative AI has reshaped many aspects of online discourse, particularly in technical forums where users seek programming-related help. Q&A platforms like Stack Overflow and GitHub Discussions have traditionally relied on user-generated content to build a repository of high-quality technical knowledge. However, with the accessibility of LLMs, users can now generate answers effortlessly, potentially altering engagement patterns and content quality. This study examines the implications of LLM-based tools on community health within these forums. Specifically, we address:

- **RQ 1**: Has the rise of LLM tools impacted how comments are received and evaluated by the Stack Overflow community?
- **RQ 2**: Are answers increasingly AI-generated, and does this impact their quality and perceived helpfulness?

Understanding these shifts is crucial for platform administrators who must devise new incentive structures and moderation strategies to maintain high-quality discussions and long-term community sustainability.

3 RELATED WORK

3.1 Relevant Background Literature

Recent studies have explored various aspects of AI-generated text detection, online community engagement, and the integration of human oversight with machine learning models. In this section, we review several key works that inform our methodology and research approach.

Voigt et al. (2017): Language from police body camera footage shows racial disparities in officer respect [5]. In this study, Voigt *et al.* analyzed police body camera footage to examine racial disparities in officer respect. Their approach combined advanced machine learning models with human verification to capture the nuances of spoken language. The study found that while automated models are capable of uncovering patterns in speech, human oversight remains crucial to interpret subtle contextual cues. This dual approach directly aligns with our methodology, where we combine GPT-Zero—a machine learning tool for detecting AI-generated text—with human verification. The police study serves as an important precedent, highlighting how the integration of both automated techniques and human judgment can lead to more robust analyses of complex speech content.

Akram (2013): An Empirical Study for AI Generated Text Detection Tools [1]. Akram's work examines the performance of various AI-generated text detection tools, including GPT-Zero. The paper reports accuracy rates for six different systems ranging from 55.29% to 97.0%, underlining the challenges inherent in distinguishing between human and machine-generated content. This study provides a foundational understanding of the performance metrics that one might expect from a tool like GPT-Zero, and it emphasizes that even the most accurate systems have limitations. These insights are critical for our research, as they justify the need for supplementing automated detection with human verification to improve reliability in our analysis of tech forum posts.

Habibzadeh (2023): GPTZero Performance in Identifying Artificial Intelligence-Generated Medical Texts: A Preliminary Study[3]. Habibzadeh evaluates GPT-Zero's ability to differentiate between human-written and AI-generated medical texts, specifically from ChatGPT. The study reports a sensitivity of 0.65, specificity of 0.90, and an overall accuracy of 0.80. Although the sample size was limited, these results indicate that GPT-Zero possesses a reasonable capability in identifying AI-generated text, albeit with room for improvement. This finding supports our strategy of integrating GPT-Zero with human oversight; by compensating for the tool's partial accuracy with expert judgment, we aim to achieve more reliable detection outcomes in our analysis of forum discussions.

Anderson et al. (2012): Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow [2]. This seminal paper investigates the dynamics of user engagement and content quality on Stack Overflow. The authors analyze patterns such as voting behavior, answer acceptance, and the longevity of questions to extract insights into what drives high-value community activity. Key contributions and methodology include:

- Analyzing large-scale Q&A data to identify engagement trends and the diffusion of knowledge.
- Using survival analysis to measure the relevance and sustained engagement of questions.
- Developing a predictive model for question longevity using logistic regression and random forest classifiers.

The relevance to our study is twofold. First, we intend to replicate and extend their engagement analysis using Stack Overflow's data dump. Second, by applying temporal segmentation, we will compare engagement metrics pre- and post-adoption of LLMs. Specifically, we will examine whether AI-generated content influences the longevity and perceived helpfulness of forum posts.

Asaduzzaman et al. (2013): Answering Questions about Unanswered Questions of Stack Overflow [4]. Asaduzzaman *et al.* focus on the factors that contribute to questions remaining unanswered on Stack Overflow. The study employs decision trees and support vector machines (SVMs) to classify unanswered questions based on features such as:

- Topic Entropy: How well a question fits into established discussion categories.
- Question Clarity: Measured through lexical complexity and readability.
- User Reputation: How quickly high-reputation users provide answers.

This work lays the foundation for understanding engagement challenges within online Q&A platforms. For our research, we plan to extend their classification model to determine if unanswered questions have become more prevalent in the post-LLM era. We will also analyze whether AI-generated questions or answers are more likely to be ignored, using techniques such as TF-IDF and coherence scoring. Additionally, we will track new user retention rates to assess if a shift toward AI-generated content impacts community engagement.

3.2 Relevance to Our Study

The reviewed literature collectively informs our methodology in several key ways:

- **Dual Approach of Machine Learning and Human Oversight:** Both Voigt *et al.* (2017) and Habibzadeh (2023) illustrate the benefits of combining automated tools (e.g., GPT-Zero) with human judgment to overcome the limitations of each method individually.
- **Engagement Metrics and Content Value:** Anderson *et al.* (2012) provide robust methodologies for assessing user engagement and content longevity, which we will adopt and extend to compare AI-generated versus human-generated content.
- **Classification and Prediction of Community Outcomes:** Insights from Asaduzzaman *et al.* (2013) regarding unanswered questions and engagement challenges will help us evaluate the impact of LLM-generated content on community responsiveness and new user retention.

These studies not only establish a strong precedent for our research but also offer concrete methodological tools that we will adapt to our specific focus on tech forum discussions. By synthesizing these studies, our research seeks to provide a comprehensive analysis of LLM-generated content's impact on tech forums, leveraging prior methodologies while introducing novel approaches tailored to the evolving landscape of AI-driven discussions.

3.3 Our Contribution to Literature

Our study builds on these prior works, and contributes to literature that focuses on the intersection of AI-generated content detection and online community dynamics. Earlier research by Voigt et al. and Habibzadeh has demonstrated the power of combining machine learning with human oversight. We strengthen literature by apply this dual approach specifically to the domain of programming question and answer forums like Stack Overflow, where the small notes of technical communication make accurate detection especially important. We also extend methodologies from Anderson et al. and Asaduzzaman et al. by explicitly comparing human and GPT-generated content across engagement metrics such as upvote distributions and top upvoted comments, which offer new insights into how LLM integration is shaping community behavior. Third, we incorporate a temporal lens, and track Stack Overflow changes from 2022 to 2024. This allows us to effectively document the rising presence of GPT-generated comments, and its evolving reception. By combining statistical modeling, detection tools, and social/technical analysis, our project contributes a new lens into AI's role in shaping conversation and trust within public, question and answer tech forums.

4 DATA

This study draws on data from Stack Overflow's public data dump (https://archive.org/details/stackexchange), and focused on posts from 2022 to 2024. ChatGPT was first introduced to the public in November of 2022, hence why these years were chosen. The datasets for RQ1 and RQ2 differ slightly, due to the nature of the analyses and associated resource constraints.

4.1 RQ 1 Data

To investigate patterns in user engagement before and after the rise of LLMs, we drew on a large dataset of Stack Overflow comments from 2022 to 2023. To prepare the data, we extracted comment entries from the complete archive by filtering for rows with timestamp metadata from the target year. The resulting dataset was split into manageable chunks of 100,000 lines to reduce the memory usage, from which we sampled evenly to create two balanced datasets. Two samples were used for different stages of the analysis. First, we constructed a dataset of 2,000 comments, 1,000 from 2022 (Pre-LLM) and 1,000 from 2023 (Post-LLM), to conduct an initial OLS regression. This smaller sample was selected for quick model testing, and included metadata such as creation date, upvote count (Score), comment text, and associated `PostId`. For the final analysis, we expanded to a larger dataset of 200,000 comments: 100,000 from the Pre-LLM era (2022) and 100,000 from the Post-LLM era (2023). This full dataset was used for both our mixed-effects regression and sentiment analysis. Comments were randomly sampled from the Stack Overflow public data dump. To prepare the data for modeling, we included several features: (1) a binary `PostLLM` variable (0 = 2022, 1 = 2023), (2) a log-transformed `CommentOrder` variable based on each comment's position in its thread, and (3) a continuous sentiment score computed using the VADER sentiment analysis tool. To account for nested structure, comments were grouped by `PostId` for use in the mixed-effects model.

4.2 RQ 2 Data

RQ2 focuses on the rise of AI-generated answers, and their impact on quality and perceived helpfulness. This research question required each post to be individually run through an AI detector (GPTZero), which was significantly more time-consuming. Due to the time intensive nature of this process, we sampled a smaller, but balanced, dataset of 1000 answers per year from 2022 to 2024, totaling to 3000 posts. For more details on this dataset limitation, please view the Limitations and Future Discussions section.

First, we filtered the relevant posts using Unix commands. For example, to isolate comments from 2022, we ran: *grep 'CreationDate="2022' Comments.xml | split -l 10000 - chunk_2022_*. This same command was then repeated for 2023 and 2024. This data was further grouped, as discussed in the later RQ 2 Methods section.

5 METHODS

5.1 RQ1 Methods

We used regression modeling to examine how comment timing, sentiment, LLM era, and author source related to community engagement, which we signify as comment upvotes (`Score`). We began with an ordinary least squares (OLS) regression on a balanced sample of 2,000 comments, then transitioned to a mixed-effects regression to account for clustering of comments by post.

5.1.1 OLS Model Setup (Balanced Sample, $N = 2000$).

We fit an OLS model using four predictors: `PostLLM`, `Sentiment`, `LogCommentOrder`, and `IsHumanWritten`. The full model specification is shown in Equation 1:

$$\begin{aligned}\text{Score} = \beta_0 + \beta_1 \cdot \text{PostLLM} + \beta_2 \cdot \text{Sentiment} \\ + \beta_3 \cdot \text{LogCommentOrder} + \beta_4 \cdot \text{IsHumanWritten} + \epsilon\end{aligned} \tag{1}$$

In this model, `Score` refers to the number of upvotes a comment received. `PostLLM` is a binary variable indicating whether the comment was posted after the release of ChatGPT (1 = post-LLM, 0 = pre-LLM). `Sentiment` is the compound sentiment score from VADER, ranging from -1 (very negative) to 1 (very positive). `LogCommentOrder` captures the log-transformed position of the comment within its post thread and reflects how early or late the comment appeared. `IsHumanWritten` is a binary indicator of whether the comment was classified as human-written (1) or AI-generated (0).

5.1.2 Post-Level Structure and Clustering.

To assess whether comments were clustered by post, we computed summary statistics on the number of comments per `PostId` in our Full Sample dataset of 200,000 comments (see Table 1). Figure 1 visualizes the distribution of comments per post.

Table 1: Summary of Comment Counts per Post

Statistic	Value
Number of unique `PostId`s	81,108
Average number of comments per post	2.47
Maximum number of comments on a post	35
Minimum number of comments on a post	1

5.1.3 Mixed Effects Model Setup (Full Sample, $N = 200{,}000$).

We used a mixed-effects regression with a random intercept for each `PostId` to better model the nested structure of comments within posts. In the mixed-effects model, the dependent variable `Score` represents the number of upvotes a comment received. The predictors `PostLLM`, `Sentiment`, and `LogCommentOrder` are the same as in our OLS model. The model includes a random intercept u_j for each `PostId` j, which controls for unobserved variation across different posts. The residual error term ϵ_{ij} captures remaining within-post variability in upvotes for each comment i. The full model is shown in Equation 2.

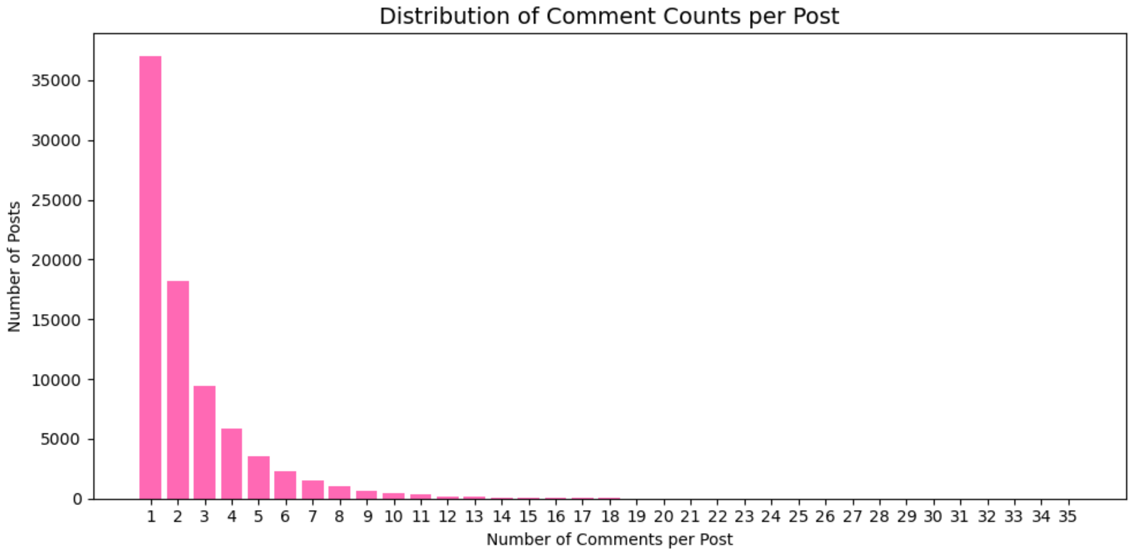


Figure 1: Distribution of the number of comments per post, based on the `PostIds` in our dataset of 200,000 comments. Most posts receive between 1 and 3 comments, but a small number of posts have 10 or more. This confirms that many comments are nested within shared posts, motivating the use of mixed-effects modeling.

$$\begin{aligned}\text{Score}_{ij} = \beta_0 + \beta_1 \cdot \text{PostLLM}_{ij} + \beta_2 \cdot \text{Sentiment}_{ij} \\ + \beta_3 \cdot \text{LogCommentOrder}_{ij} + u_j + \epsilon_{ij}\end{aligned} \tag{2}$$

5.1.4 Sentiment Analysis Procedure. We applied the VADER sentiment analysis tool to classify comment text as *positive*, *neutral*, or *negative* and used the compound score thresholds: scores ≥ 0.05 were labeled positive, scores ≤ -0.05 were negative, and all others were neutral. This procedure was applied to the full sample of 200,000 comments. A Chi-square test was used to compare sentiment distributions across years.

5.2 RQ 2 Methods

RQ 2 sought to understand whether answers on Stack Overflow are increasingly AI-generated and if this affects their quality and perceived helpfulness. To do this, Eisha utilized a multi-step approach; this approach had two phases: data processing and statistical testing. The goal was to analyze trends in GPT-generated comments over the eyars, and assess whether they differ significantly in terms of upvotes compared to human authored comments.

- **Data Collection** A Selenium script was written to automate the process of scraping data from a set of forums for each year (2022, 2023, and 2024). The script processed each comment on posts, and piped the comment into GPTZero, which determined whether the comment was GPT-generated.The result for each comment (GPT-generated score) was written to a text file. This score ultimately acted as a new "GPTScore" column within the data for each comment.
- **Data Preprocessing** For each comment from each year, the scraped comment data was parsed, and relevant attributes were extracted: PostId, CommentId, Score, and GPTScore. These were stored in a structured format for further analysis.
- **Data Aggregation** For each post, the highest-scoring comment was selected, which could be either GPT-generated or human written; note that if a comments GPTScore was greater than zero, then it qualified as GPT-generated. The data was grouped by PostID to ensure only the most relevant comment for each post was considered.
- **Statistical Analysis**
  - **Linear Regression** A linear regression analysis was done to test if there is a significant trend in the percentage of GPT-generated comments across the years 2022 to 2024.
    - **Null Hypothesis** There is no significant trend in the percentage of GPT-generated comments over the years; the proportion of GPT comments remains constant).
    - **Alternative Hypothesis** There is a significant trend in the percentage of GPT-generated comments over the years: the proportion of GPT comments either increases or decreases.

    The trend was evaluated using the slope of the regression line and the p-value of the test. If the p-value for the slope was less than 0.05, the null hypothesis would be rejected. This would indicate a statistically significant trend in the percentage of GPT-generated comments over time.
  - **Binomial Test** The binomial test was conducted to determine whether GPT-generated comments on Stack Overflow posts from 2022-2024 are significantly less likely to receive more than 2 upvotes. The threshold of 2 upvotes was chosen as a benchmark for positive engagement. In many online forums, comments that receive more than 2 upvotes are often

considered to have attracted an acceptable level of community approval, beyond the default visibility or random exposure. This threshold helps distinguish between comments that just appear on a post and work for one person, and those that users actively found useful or informative.

  - ∗ **Null Hypothesis** The probability of a GPT-generated comment receiving more than 2 upvotes is 20% or greater
  - ∗ **Alternative Hypothesis** The probability of a GPT-generated comment receiving more than 2 upvotes is less than 20%

The hypothesis was tested using the cumulative distribution function (CDF) of the binomial distribution. If the p-value from the test was below 0.05, then the null hypothesis would be rejected. This would indicate that GPT-generated comments are significantly less likely to receive more than 2 upvotes, and tend to not meet the threshold of "community acceptance and approval".

- – **Evaluation** The results of the linear regression and binomial tests were analyzed to evaluate the hypotheses regarding the behavior of GPT-generated comments across years, as well as how they compare to human written ones.
- – **Tools** Beyond the data and computational hardware required to perform this research, the primary tools used for analysis included Python libraries, such as matplotlib for data visualization and scipy.stats for statistical tests (linear regression and binomial).

6 RESULTS

6.1 RQ1 Results

6.1.1 OLS Regression (Balanced Sample). The OLS model on 2,000 comments was not statistically significant overall ($F(4, 1995) = 1.10$, $p = 0.357$), explaining only 0.2% of the variance in comment upvotes ($R^2 = 0.002$). Post-LLM comments received significantly fewer upvotes on average ($\beta = -0.066$, $p = 0.042$), but other predictors were not significant (see Table 2).

Table 2: OLS Regression Predicting Upvotes (Balanced Sample, $N = 2000$)

Variable	Coefficient	Std. Error	p-value
Intercept	0.347^{**}	0.106	0.001
PostLLM	-0.066^{*}	0.032	0.042
Sentiment	0.002	0.041	0.953
LogCommentOrder	-0.022	0.291	0.939
IsHumanWritten	-0.049	0.106	0.642

Note: $^{*}p < 0.05$, $^{**}p < 0.01$.

6.1.2 Mixed Effects Regression Results. The mixed-effects model on 200,000 comments revealed that: (i) Post-LLM comments received fewer upvotes ($\beta = -0.034$, $p < .001$), (ii) Later comments were less upvoted ($\beta = -0.117$, $p < .001$), (iii) More positive sentiment slightly predicted fewer upvotes ($\beta = -0.029$, $p < .001$), and (iv) Post-level variance was estimated at 0.142 (see Table 3), indicating that differences between posts explained a non-trivial portion of the variation in upvotes. While statistically significant, these effects were modest in size, suggesting that other unmeasured factors such as comment content or user reputation may play a larger role in determining engagement.

Table 3: Mixed-Effects Regression Predicting Comment Upvotes

Variable	Coefficient	Std. Error	p-value
Intercept	0.355^{***}	0.003	<0.001
PostLLM	-0.034^{***}	0.004	<0.001
Sentiment	-0.029^{***}	0.004	<0.001
LogCommentOrder	-0.117^{***}	0.003	<0.001
Random Effects:			
PostId Variance	0.142	—	—

Note: $N = 200{,}000$ comments across 81,108 posts.
$^{***}p < 0.001$.

6.1.3 Sentiment Analysis Results. A Chi-square test revealed a statistically significant shift in sentiment between eras, $\chi^2(2) = 28.74$, $p < .001$, but effect sizes were minimal. Positive sentiment decreased from 49.71% to 49.00%, neutral sentiment dropped from 30.38% to 30.12%, and negative sentiment rose from 19.92% to 20.88% (see Figure 2).

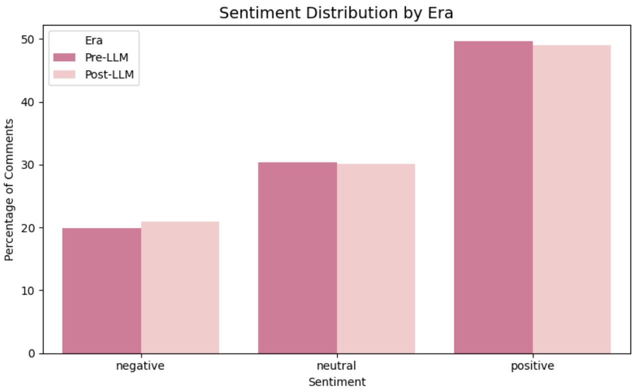


Figure 2: Sentiment distribution of Stack Overflow comments by era. While a Chi-square test indicates a statistically significant difference ($\chi^2(2) = 28.74$, $p < .001$), the overall sentiment distributions are largely similar between the Pre-LLM and Post-LLM periods.

6.2 RQ2 Results

There were strong insights across both the linear regression and binomial test analyses, which reveal trends and differences in the prevalence and engagement of GPT-generated comments on Stack Overflow. The analyses not only confirm that GPT-generated comments are becoming increasingly common over the years, but also uncover a consistent pattern of community engagement for these comments.

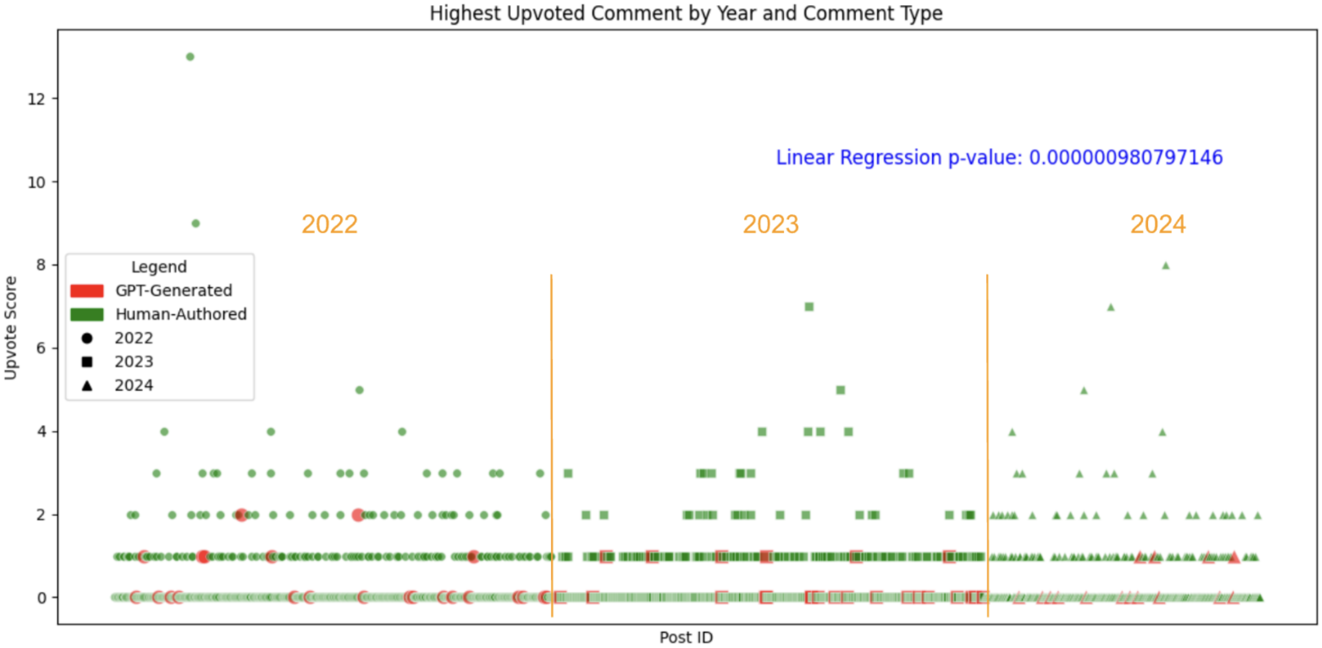
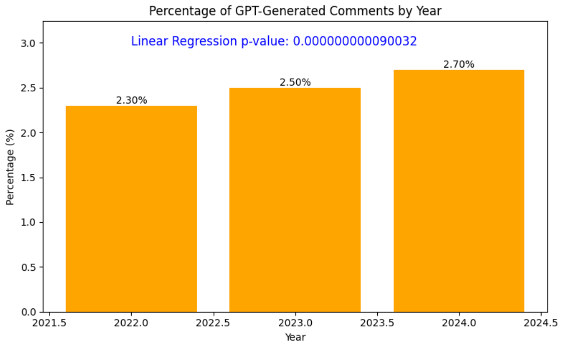


Figure 3: Highest upvoted comment by year and comment type

- **Linear Regression** The linear regression analysis of the percentage of GPT-generated comments over the years resulted in a p-value of 9.00×10^{-11}. This p-value is also below the 0.05 threshold, so we reject the null hypothesis. We then accept the alternative hypothesis, that there is a significant trend in the increase of GPT-generated comments across the years of 2022 - 2024.



- **Binomial Test** The binomial test for GPT-generated comments receiving more than 2 upvotes yielded a p-value of 9.81×10^{-7}, which is far below the 0.05 significance threshold. As a result, we reject the null hypothesis that 20% of GPT-generated comments would receive more than 2 upvotes. We accept the alternative hypothesis, that GPT-generated comments are significantly less likely to receive more than 2 upvotes. See Figure 3.

The findings strongly suggest that the proportion of GPT-generated comments on StackOverflow has been steadily increasing from 2022 through 2024, but they still represent a minority of total comments. This indicates a broader trend in the integration of AI-generated content into online forums. This rise may reflect the growing accessibility and adoption of large language models by users seeking quick or automated contributions to forum discussions. However, the binomial test results highlight a critical limitation: GPT-generated comments are statistically significantly less likely to receive more than 2 upvotes, which is our defined benchmark for minimal community endorsement or approval. This reflects a deeper issue of community trust and perceived value and helpfulness of a GPT-generated comment. Even in cases where GPT-generated responses surfaced as the top comment (by upvote score), they often failed to attract the level of positive engagement that typically occurs for valued contributions. In contrast, human-authored comments more frequently surpassed this threshold, and often receiving considerably more upvotes, as shown in the graph. This suggests that such comments may reach the top by default, rather than by earning merit from the community. The community's hesitance to upvote GPT-generated content beyond a basic 2 upvote threshold may reflect skepticism about the originality, depth, or contextual understanding of these comments. These results together suggest that while GPT-generated content is becoming more common, it is not yet perceived as a trusted voice within online communities.

7 DISCUSSION

7.1 Summary of Findings

7.1.1 ***RQ1:*** The OLS regression model was not statistically significant overall (p = 0.357), suggesting that the set of predictors did not explain a significant amount of variance in upvote counts ($R^2 = 0.002$). The result of RQ1 indicates that the rise of LLM tools has influenced how comments are received on Stack Overflow. Specifically, post-LLM comments received fewer upvotes on average, suggesting that there exists a potential shift in how the community evaluates contributions in the post-LLM era. However, the OLS regression model explained very little variance in upvotes, and other factors such as `Sentiment`, `LogCommentOrder`, and `IsHumanWritten`, were not significant predictors. This implies that while timing relative to LLM adoption matters to a small extent, community engagement is largely shaped by factors that are not captured in the current model. The mixed-effects model indicates that comments posted after the LLMs were introduced tend to receive fewer upvotes, which suggests a subtle shift in user engagement pattern in the post-LLM era. In addition, comments that were posted later in threads and those comments with more positive sentiment were also associated with reduced upvotes. Despite these statistically significant effects, their small magnitudes (as shown in Table 3), along with relatively large post-level variance, indicate that comment reception varies meaningfully depending on the specific post context. Although a Chi-square test detected a significant shift in sentiment distribution between the Pre-LLM and Post-LLM eras, the changes were minimal in practical terms. The proportion of sentiment categories changed slightly, indicating overall stability in sentiment of comments has not meaningfully changed with the rise of LLM tools. This indicates that while LLM adoption may influence how comments are received, it has not substantially impacted the sentiment of the comments themselves.

7.1.2 ***RQ2:*** The results indicate an increasing presence of GPT-generated content on Stack Overflow. While some GPT-generated comments appear as top-ranked responses, they consistently fail to receive substantial community upvotes. This suggests that although the use of AI-generated responses are increasing, it receives less upvotes, suggesting that it failed to attract positive engagement by users, which are less valued by the users. Human-written comments continue to attract stronger engagement, which indicates that users still prioritize perceived authenticity and expertise in evaluating comment quality.

7.2 Implications

This study has multiple implications for the future of online knowledge-sharing platforms in this new era of LLMs.

The most practical implications is that for community moderators or platform designers, the increasing presence of AI-generated content poses many challenges and opportunities. While LLMs can fill knowledge gaps and speed up responses to questions, they also risk flooding forums with low-quality responses or misleading information.

As such, our research suggests a design implication that with the growing presence of AI-generated content on these technical platforms, platform moderators or engineers might want to consider implementing methods to identify and flag AI-generated posts and comments, so that users continue to enagage and trust the platform community for providing quality knowledge. This could include optional labels for AI-assisted posts, or UI nudges that encourage users to review or edit their AI-generated content. These designs could ultimately empower users to better interpret the reliability of the content they consume.

From a policy standpoint, a key implication is that platforms must tread very carefully in terms of how to treat AI-generated content. Too strict of policies might unfairly punish users who use LLMs responsibly, or whose writing style coincidentally resembles AI output. On the other hand, too relaxed of policies could lead to the erosion of trust in these platforms, which rely on community sourced knowledge. Our findings support the need for transparent AI content policies, which balance accountability and fairness.

Ultimately, these study highlights that the integration of LLMs into public knowledge sharing ecosystems is not just a technical challenge, but also a social one too; it will require careful consideration of user behavior, community norms, platform design, and governance structures.

7.3 Ethical Considerations

This study poses some ethical concerns, specifically around the usage of AI detection tools, as well as the consequences of misclassifications. Within our study, the most prominent bias in our analysis stems from the AI detectors themselves. Many studies show that the models backing these AI detectors often have false positives and false negatives; this is because of the fact that AI detection models rely on lingustic style, sentence structure, or even fluency to denote if something is AI-generated or not. For example, if a Stack Overflow user writes in formal, structured English, we might inaccurately flag their writing as AI-generated, with our current study methods. At the end of the day, this mislabeling could unfairly affect this users credibility within the Stack Overflow community.

Another thing to keep in mind is the application of such studies, or even AI detection tools, by forum moderators or automated systems. If this knowledge is taken too far, it might result in disproportionate consequences. For example, a platform like Stack Overflow might find our study valuable, as they want to mitigate "bot" activity on their site to keep quality and engagement. However, if Stack Overflow were to use such flagging methodologies to ban users posting suspected AI content, then they risk punishing people who are genuinely contributing well to the platform. This could also effect other user participation, as people might fear to take part in conversation since they are worried about if their post could get them banned. In the long term, these tactics could truly erode trust in community.

Overall, while our study aims to analyze these trends in AI content on such platforms, any interventions based on this research paper should be approached with caution.

7.4 Limitations and Future Discussions

7.4.1 ***Limitations.*** While this study offers many valuable insights into the evolving dynamic of LLM tools on technical question and answer forums, there were also several limitations which constrained the interpretation, scale, and generalization of our study.

- **AI-Detection Method**: The first limitation, as discussed above, was the fact that detection of AI-generated content is inherently biased and imperfect. Although tools like GPTZero are useful, they rely on a variety of linguistic patterns that do not always successfully distinguish between human and AI writing. Posts that are co-authored by humans and LLMs or paraphrased from LLM suggestions pose many challenges for binary classification of if a post was AI-generated or not. We foresee this to be an increasing challenge, especially as LLMs get to be more and more familiar with human writing styles. Ultimately, these ambiguities may lead to under-counting or overestimating the presence of AI-generated content in our data.

- **Large-Scale Detection:** A large logistical constraint faced during this process was in running the GPTZero detection on large amounts of data. As students, we did not have access to paid API keys for LLM detection tools. To work around this, one of the authors (Eisha Peyyeti) developed a custom Selenium script to automate the submission of posts to the GPTZero web interface. While this allowed us to do batch processing of posts without paying, it also introduced a significant time constraint, and limited the amount of posts we could analyze. For example, running just 1000 posts through the Selenium script took around two to three hours alone, and caused Eisha's laptop to overheat and utilize much of the CPU. Repeating this process on larger datasets would have been good for our statistical analysis, but it simply was not feasible. Consequently, our data set was diverse, but ultimately smaller than ideal for high-confidence statistical inference (particularly in regards to RQ 2).
- **Limited Platform:** A final limitation was that our study was constrained to just the Stack Overflow platform, which is a platform focused heavily on code and programming questions. The results of our study may not generalize to other question and answer platforms that are less code heavy and have different dynamics, like Quora or Reddit. Moreover, the Stack Overflow user base tends to be more technically saavy, which may influence how frequently and effectively users utilize LLM tools on such platforms. On one hand, on more technical platforms, this could mean higher adoption rates of LLMs due to familiarity with the technology; on the other hand, this could also mean that users may be more skeptical of AI-generated content, and don't use it as much. In contrast, on less technical platforms, users may engage with LLMs differently, or not at all. Overall, these variations highlight a need for broader studies across diverse communities and subject areas.
- **RQ1 Limitations:** First, while our RQ1 analysis uses a large dataset of 200,000 comments, this scale comes with limitations. The large sample size increases statistical power, which may lead to the detection of statistically significant effects that are not meaningful. For example, predictors with very small effect sizes yielded highly significant p-values, which should be interpreted with caution. Second, although the mixed-effects model accounts for post-level clustering, we do not control for other potentially confounding variables such as comment length, user reputation, or question complexity. Lastly, the use of upvotes as a proxy for helpfulness assumes consistent voting behavior across users and time, which may not hold given platform norms and dynamics.

7.4.2 ***Future Discussions.*** Looking ahead, there are several directions for the continuation of this study. We see that some kind of longitudinal study could be done to track how individual users and their communities adapt to the rise of AI assistance; this could include whether or not norms shift, if moderation evolves, if human-AI collaboration becomes normalized, or even experimental interventions, like labeling supposed AI content with a tag on platforms. The latter views could help assess how disclosure of AI-generated content affects engagement and trust. Finally, we envision that future collaborations with platform developers could allow for more robust AI-detection and feedback mechanisms, which can ultimately balance innovation with user safety and fairness.

8 CONTRIBUTIONS

8.1 Eisha Peyyeti

- **Code and Data Contributions:**
  - **Ideation** During the ideation phase, I contributed to refining the scope and direction of the research, particularly around the use of GPT-generated content and validation of this GPT-detection on tech forums. I suggested integrating GPT-Zero, an AI text detection tool, to analyze Stack Overflow posts, and brought in ideas of human-validation from the "Language from police body camera footage shows racial disparities in officer respect" research paper.
  - **Formulation and Execution** I supported the initial data collection and chunking, as well as independently answered the entirety of RQ 2, from the initial code, to the final analysis and report. To handle the large dataset, I wrote a series of bash commands to efficiently chunk the data into manageable parts. I also wrote multiple pieces of regex code to accurately parse out comment text from other text and headers within the file. Once the data was chunked and properly prepared, I shifted my focus to automating the detection of GPT-generated comments. I learned how to use Selenium, and wrote a Selenium script to interact with the dataset and systematically run every comment of our data through GPTZero. The script processed each comment, passed it through GPTZero to determine if it was likely AI-generated, and then recorded a GPT score for each comment based on the likelihood that it was generated by an AI. This score was then saved in a new "GPTScore" column in the dataset. After generating the GPT score for each post, I compiled a Google sheet containing a random sample of 100 posts for each year (2022, 2023, and 2024). Each post in this dataset was paired with its respective GPT score, which allowed my teammates to perform human validation. After this, I wrote the code and performed an analysis to answer RQ 2 effectively. I wrote a plethora of code to run several statistical analyses to explore the relationship between GPT-generated posts and user engagement. I conducted the binomial test to examine whether GPT-generated comments were statistically less likely to receive more than 2 upvotes, which a key metric for community engagement. I also conducted linear regression to explore the trend in GPT-generated comments over time (2022 to 2024) and how this related to upvotes and other engagement metrics. For the binomial test, I put together both the null and alternative hypotheses, and considered the distribution of upvotes for GPT-generated content. In the linear regression analysis, I examined the trend of GPT-generated comments over time, and also put together the null and alternative hypothesis.

  - **Presentation** Once the analyses were complete, I compiled the results and plotted all results through visual means (which are the two graphs observed in the RQ 2 results section). I summarized these results throught the paper. More specific paper contributions of mine can be seen in the "Final Paper Contributions" below. Throughout the process, I also assisted other team members as needed.
- **Final Paper Contributions:** I independently wrote the entirety of the following sections of the final paper: Abstract, Our Contribution to Literature, RQ 2 Data, RQ 2 Methods, RQ 2 Results, Implications, Ethical Considerations, and Limitations and Future Discussions. I also contributed heavily to the Related Work, and wrote on the three following papers: Voigt et al. (2017): Language from police body camera footage shows racial disparities in officer respect, Akram (2013): An Empirical Study for AI Generated Text Detection Tools, and Habibzadeh (2023): GPTZero Performance in Identifying Artificial Intelligence-Generated Medical Texts. Along with this, throughout the paper, I updated initial written content to accurately reflect our finalized methodology and findings.

8.2 Sunwoo Baek

I was responsible for preparing the datasets used across both phrases of our analysis. I wrote the shell scripts and python code that are necessary to extract, filter, and sample comments from the Stack Overflow data archive, ensuring that we had balanced and have representative samples from 2022 and 2023. To handle large size of the data, I implemented a chunking system that split the raw XML data into manageable blocks of 100,000 lines, and then evenly sampled across these chunks to generate clean datasets of 100,000 comments per year. For RQ1 and RQ2, I designed and implemented the comment time offsets by grouping comments by `PostId` and measuring the time difference from the first comment in each thread. I visualized the relationship between time offset and upvote score. I also handled the fuzzy matching between helpful comment samples and GPT-detection outputs, using RapidFuzz to accurately align comments with associated GPT scores. Furthermore, I debugged data inconsistencies, verified text normalization, and supported teammates in troubleshooting script execution. I worked on RQ1 Data, RQ1 Methods, and Summary of Finding sections of the paper.

8.3 Bridget Agyare

I helped shape the project scope by giving feedback on the feasibility of different research directions and suggesting more manageable framings of our research ideas. I also led discussions on which methods to use for RQ1 and landed on sentiment analysis and regression modeling as our primary approaches based on istructor feedback. I implemented the RQ1 analyses presented in the paper, including sentiment classification using VADER, OLS and mixed-effects regression modeling, Chi-square tests, and the post-level clustering analysis. I also generated the visualizations and regression tables for RQ1, and wrote the corresponding Data, Methods, Results, and Limitations sections. I lightly edited other sections of the paper to ensure cohesion. Throughout the project, I helped refine our interpretation of the findings and ensured that the analyses remained aligned with our research questions.

8.4 Himnish Jain

I initiated the original core research idea for this project, proposing the exploration of how Large Language Models like ChatGPT are influencing user engagement and content quality on Stack Overflow. I helped define the research questions and shaped the project's initial scope, particularly emphasizing the dual focus on detection and community impact. I helped define the research questions and conceptual scope of the project, especially RQ1, which focuses on changes in community reception of comments before and after the emergence of LLMs. I conducted the background literature review, synthesizing prior studies on AI-generated content detection and online engagement patterns, including key works by Habibzadeh et al., Anderson et al., and Asaduzzaman et al. I also carried out a comparative analysis using code to examine why GPTZero flagged certain comments as AI-generated, even when written by humans, and explored stylistic differences that may have contributed to misclassification. This analysis offered valuable insight into the limitations of AI detection tools and informed our discussion on model bias and false positives, along with the rise of AI-like content. Additionally, I contributed to interpreting the results for RQ1 and provided feedback throughout the writing and analysis process to ensure alignment between our methods, results, and overall research goals.

REFERENCES

[1] Akram, A. An empirical study of ai generated text detection tools. *Opast Group LLC* (2023).

[2] Anderson Ashton, Daniel Huttenlocher, J. K., and Leskovec, J. Discovering value from community activity on focused question answering sites: a case study of stack overflow. *Association for Computing Machinery* (2012), 850–858.

[3] Habibzadeh, F. Gptzero performance in identifying artificial intelligence-generated medical texts: A preliminary study. *Journal of Korean Medical Science of the National Academy of Sciences 38* (Sept. 2023), 6521–6526.

[4] Muhammad Asaduzzaman, Ahmed Shah Mashiyat, C. K. R., and Schneider, K. A. Answering questions about unanswered questions of stack overflow. *Institute of Electrical and Electronics Engineers* (2013), 97–100.

[5] Rob Voigt, Nicholas P. Camp, V. P. W. L. H. R. C. H. C. M. G. D. J. D. J., and Eberhardt, J. L. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences 114* (2017), 6521–6526.