# Investigating the Relative Impact of Generative AI vs. Humans on Voluntary Knowledge Contributions

Guohou Shan
Temple University
guohou.shan@temple.edu

Dan Pienta
University of Tennessee
dpienta@utk.edu

Jason Bennet Thatcher
Temple University
jason.thatcher@temple.edu

## Abstract

*Voluntary knowledge contributions on question and answer (Q&A) platforms are important for users, platforms, and organizations. Generative Artificial Intelligence (GAI) techniques have made it possible to automatically generate voluntary knowledge on Q&A platforms. The relative impact of GAI vs. humans on users' voluntary contribution of knowledge to Q&A platforms has yet to be explored. On the one hand, GAI can generate highly accurate answers because it is trained on large volumes of diverse, high-quality data. On the other hand, GAI can produce incorrect answers and fabricated facts. Using data from one of the largest Q&A platforms, Stack Overflow, we apply fixed effects models to understand the relative impact of GAI vs. human contributors on answer quality. We find that, on average, GAI answers receive lower scores and are shorter, but they can also be easier to read, more positive, and more objective. Our study has both theoretical and practical implications.*

**Keywords:** Knowledge contribution, generative AI vs. human, answer quality, reputation, fixed effects.

## 1. Introduction

Voluntary knowledge contribution on Q&A platforms has become increasingly important for users, platforms, and organizations. Users who contribute knowledge can reap rewards, including potentially opening up new job opportunities or leading to promotions within their organization (Malgonde et al. 2023). Users' voluntary knowledge contribution also plays a critical role in sustaining the development of platforms (Wang et al. 2022a). For organizations, voluntary user knowledge contribution can improve productivity (Huang et al. 2022).

In recent years, significant advances in generative artificial intelligence (GAI) techniques have enabled the generation of voluntary knowledge through automation. One such innovation is the GAI chatbot, ChatGPT, which is based on large language models such as GPT-3.5 and incorporates reinforcement learning from human feedback. Developed by OpenAI and released on November 30, 2022, ChatGPT is capable of mimicking human users in answering questions in a conversational manner (OpenAI 2022). However, a Q&A knowledge platforms industry leader, Stack Overflow (SO), banned the use of ChatGPT on December 5, 2022 (Stackoverflow 2022).

However, how the voluntary knowledge contributed by generative AI vs. human users is different on Q&A platforms has yet to be explored. On the one hand, GAI has the potential to generate highly accurate answers. Models such as ChatGPT are trained on diverse and high-quality data. These models have been shown to provide organized, logical, and detailed answers to questions (Guo et al. 2023). In addition, Jia et al. (2023) found that AI can improve an individual's ability to answer questions, particularly by handling repetitive questions, allowing humans to focus on generating more creative responses.

On the other hand, ChatGPT has been shown to occasionally produce incorrect answers and fabricated facts (Borji 2023). Understanding the potential relative quality of voluntary knowledge contributed by GAI and human users on Q&A platforms is crucial because it has implications for users, platforms, and organizations (e.g., Huang et al. 2022; Wang et al. 2022a). For example, from a practical point of view, without voluntary knowledge contributions from users, knowledge platforms such as YouTube and Stack Overflow can not sustain (Wang et al. 2022a). Therefore, we address the following research question.

**RQ1:** *How do generative AI contributions differ from human contributions on Q&A platforms?*

Answering this question has become very important. Q&A voluntary knowledge platforms have implemented policies to prohibit the use of GAI in answering questions on their platforms. For example, Stack Overflow (SO), a major Q&A platform, banned the use of GAI on December 5, 2022 (Stackoverflow 2022). However, it is unclear whether such bans are necessary because we do not know if GAI can contribute as well as humans to Q&A platforms. By addressing RQ1, we can provide guidance on the management and implementation of policies regarding the use of GAI on voluntary knowledge platforms, with the goal of improving platforms' evolution.

Our study shows that the answers generated by GAI and humans differ on several quality measures. Specifically, GAI responses have lower answer scores (i.e., upvotes) with shorter and easier-to-read content, but they can be more positive and objective. Furthermore, it is currently unclear whether GAI-generated responses differ from those of users with higher community reputations. Previous studies (e.g., Chen et al. 2018; Xu et al. 2020) have shown that users' motivation to contribute to Q&A platforms may be due to peer recognition, as voluntary knowledge contribution is prosocial behavior. A higher community reputation indicates that a user's contributed content is more liked and recognized by other users (Burtch et al. 2022). For those users, they may form a regular habit of contributing high-quality content (Kim et al. 2005; Bateman et al. 2011), which can make their generated answers furtherly different from GAI. To understand how user reputation may affect the relative impact of GAI on humans' voluntary knowledge contributions, we propose our second research question as follows:

**RQ2:** *How does users' reputation moderate the relative impact of generative AI and humans on voluntary knowledge contribution?*

Answering RQ2, we provide knowledge platform managers with insights on when to use GAI to improve voluntary knowledge contributions on their platform. Specifically, we find that user reputation can enhance the relative impact of GAI and human users on answer scores and length while weakening the relative impact on sentiment and subjectivity. However, the impact on answer readability is not affected.

This research contributes in two ways. Theoretically, we contribute to the understanding of voluntary knowledge contribution in the information systems (IS) literature (e.g., Wang et al. 2022a) and the relative impact of AI vs. humans in the IS literature (e.g., Wang et al. 2022b) by investigating their impacts on voluntary knowledge contribution on Q&A platforms. Practically, this research provides managerial implications for platform managers to create policies that leverage GAI on their platform and inform users and organizations to understand the relative quality differences of knowledge contributed by GAI and humans.

# 2. Theoretical background and research hypotheses

In this section, we propose our research model (shown in Figure 1) and research hypotheses.

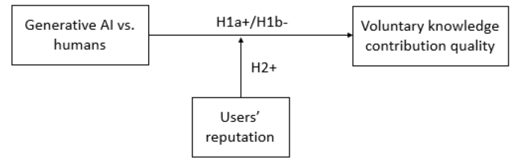


**Figure 1. Research model.**

## 2.1. Competing mechanisms of the relative impact of generative AI and humans

Recently, AI has advanced at a very fast pace. Luminaries like Bill Gates believe that the era of AI has begun (Gates 2023). Prior research also shows that AI works well in many fields, including sales (Luo et al. 2021), investment (Ge et al. 2021), live-streaming e-commerce (Wang et al. 2022b), and online review generation (Shan & Jia 2024).

GAI, such as ChatGPT, are trained on large language models, incorporating reinforcement learning from human feedback (e.g., OpenAI 2022). GAI is capable of answering questions in an organized and logical manner at high speed (Guo et al. 2023). Given the vast amount of knowledge it is trained on, GAI holds the potential to generate higher quality knowledge than a human.

However, GAIs such as ChatGPT have been shown to occasionally produce incorrect answers and fabricate facts (Borji 2023), which affects the quality of the generated answers. In addition, the quality of GAI output depends on the quality of the prompts it receives (White et al. 2023). Thus, the quality of GAI's knowledge contributions may be inferior to that of an average human. Thus, we hypothesize that:

**H1a:** *Voluntary knowledge contribution from GAI has higher quality than from human users.*

**H1b:** *Voluntary knowledge contribution from GAI has lower quality than from human users.*

## 2.2. Moderating impact of users' reputation

Previous studies have established that intrinsic motivation, such as users' sense of reciprocity, recognition, and self-image, has been shown to significantly influence their engagement in prosocial behaviors such as voluntary knowledge contribution (e.g., Xu et al. 2020). In this study, we use reputation as a proxy for users' motivation to realize recognition and enhance their self-image. We posit that users' community reputation may influence the quality of their *voluntary* knowledge contributions because gaining reputation is one of the main motivations for users to engage in voluntary contribution reputation (Chen et al. 2018). In addition, users' community reputation measured by peer recognition signals their contribution experience because a higher reputation indicates a high experience of contributing high-quality content. Literature shows that users with more experience in contributing high-quality content can develop a regular habit of voluntarily contributing high-quality content (Kim et al. 2005), which makes differences between responses generated by GAI and humans even more prominent. Therefore, community reputation may increase the relative influence of AI and humans on voluntary knowledge contribution. Thus, we hypothesize that:

**H2:** *The relative impact of generative AI and humans on voluntary knowledge contribution is more salient for users with higher community reputations.*

# 3. Research context and data

## 3.1. Research context

We use SO as the research context. SO is a question-and-answer (Q&A) platform for developers that was launched in 2008. It has 14 million registered users and 100 million unique visitors per month in 2022 (David 2023). Registered users can voluntarily ask and answer questions on the platform, resulting in a total of 23 million questions and 34 million answers, with four questions asked per minute (David 2023).

Since the launch of ChatGPT by OpenAI on November 30, 2022, many SO users have used it to answer questions. For example, Andrew Shearer shared his experience using ChatGPT to answer questions about SO within the developer community (Shearer 2022). As a result, on December 5, 2022, SO implemented a policy prohibiting the use of ChatGPT (Stackoverflow 2022). The introduction of ChatGPT provides an opportunity to investigate the relative impact of GAI and humans on voluntary knowledge contribution. We identify ChatGPT-generated responses as GAI responses (treatment group) and responses to the same questions not generated by GAI as human responses (control group). Having both treatment and control groups allows us to empirically investigate the relative influence of GAI and humans on users' voluntary contribution.

## 3.2. Treatment identification

To investigate the relative impact of GAI and humans, we need to identify the processed responses generated by GAI. Several tools are available to detect whether textual content is generated by a GAI, such as GPTZeroX developed by Edward Tian (Princeton), AI Text Classifier and GPT-2 Output Detector developed by OpenAI, and DetectGPT developed by Stanford University (Alcántara 2023; Bowman et al. 2023; Kirchner et al. 2023). These methods have advantages and disadvantages. For example, GPTZeroX, OpenAI GPT-2 Output Detector, and DetectGPT can detect if text is generated by AI with shorter content, but they also have minimum word length requirements (e.g., 50 tokens, 250 characters, and 48 tokens, respectively). The OpenAI text classifier can recognize text generated by GPT3.5; however, it will be very unreliable if the text contains less than 1,000 characters. In this study, we use OpenAI GPT-2 Output Detector to identify the responses using GAI-ChatGPT because the text on SO is usually short (our data has an average of 27 words), and its detection outputs are perceived as convincing and can have 99.3% accuracy (Solaiman et al. 2019).

## 3.3. Data

To investigate our research questions and model, we collected data from the SO platform during the period from the launch of ChatGPT on November 30, 2022, until December 4, 2022 (it was banned on December 5, 2022). The collected dataset contains 17,916 questions and 60,668 answers from 14,534 users. Using our strategy to identify answers generated by GAI, we find that there are 16,715 answers generated by GAI for 3,791 questions. We also include the answers to the same questions as a control group. Our final sample consists of 40,023 responses.

We focus on investigating the relative impact of GAI and humans on knowledge contribution quality. Following Shan and Qiu (2023), we measure the quality from two dimensions: the community acceptance of the answer (answer score) and the textual quality of the answer. We measure community acceptance using the number of upvotes received for the answer, which represents the quality of the answer as rated by other users (Lee et al. 2019). To measure the textual quality of the post, we use the length, readability, sentiment, and subjectivity of the answer, which are commonly used in previous research, to indicate the quality of knowledge posts (e.g., Qiao et al. 2020). We also include control variables for user question variants, including tenure and morning, because tenure signals answering habits (Chen et al. 2018) and affects user contribution (Shan et al. 2022), and timing indicates the cognitive load of answering (Guo et al. 2022). Thus, our data consists of the variables listed in Table 1. To gain a better understanding of our data, we provide data summary statistics in Table 2 and perform a correlation analysis of the variables, where we found that only answer length and readability have a high correlation (r = 0.74), while all other variables have negligible correlations. Hence, we control readability for investigating the impact on length and length for investigating the impact on readability.

**Table 1. Variables and Definitions**

| Category | Variables | Definitions |
|---|---|---|
| *Independent Variable* | $GAI_i$ | It shows whether answer *i* is generated by generative AI. |
| *Dependent Variables* | $Score_i$ | It represents the community acceptance measured by the received upvotes of answer *i*. |
| | $Length_i$ | It represents the length in terms of the number of words of answer *i*. |
| | $GF_i$ | It represents the readability of answers *i*. We used gunning-fog measure, following Khern-am-nuai et al. (2018). It measures the years of formal education needed to understand the generated answers. |
| | $Sentiment_i$ | It represents the sentiment score of answers *i*, indicating whether the answer has positive or negative tones. We use Python NLTK and lexicon, following Shan et al. (2018) and Shan et al. (2021). |
| | $Subjectivity_i$ | It represents the subjectivity of answers *i*, indicating whether the content of the answer is subjective or objective. We use Python TextBlob, following Mousavi et al. (2020). |
| *Moderator and Control Variables* | $Reputation_j$ | It represents the community reputation points of user *j*. We measure it by differencing gained reputation points from lost reputation points (Cooper 2020). Behaviors like posting questions, answers, and articles can receive reputation points (e.g., 10 points each) if they are voted up while being voted down decreases reputation points (e.g., 2 points each). |
| | $Tenure_{ij}$ | It represents the tenure of the user *j* when they generate answer *i* on the SO platform. We measure it in terms of days. |
| | $Morning_{ijk}$ | It represents whether the answer *i* from user *j* is generated in the morning. |

**Note.** The unit of analysis of this study is an answer level. In other words, all our variables are measured at the answer level in our sample.

**Table 2. Variable Summary Statistics**

| Variables | Mean | Std.Dev. | Min | Max |
|---|---|---|---|---|
| Score (1) | .23 | .62 | 0 | 13 |
| Length (2) | 27.6 | 22.22 | 1 | 450 |
| GF (3) | 10.37 | 8.37 | 0 | 360.4 |
| Sentiment (4) | .001 | .1 | -1.19 | 1.24 |
| Subjectivity (5) | .34 | .30 | 0 | 1 |
| GAI (6) | .42 | .49 | 0 | 1 |
| Reputation (7) | 39491 | 109425 | 1 | 1371965 |
| Tenure (8) | 2210.4 | 1638.87 | 1 | 5240 |
| Morning (9) | .41 | .49 | 0 | 1 |

From Table 2, the mean of the treatment is 0.42, which means that 42% of the observations (answers) are generated by using GAI. The average score, length, readability, sentiment, and subjectivity of an answer are 0.23, 27.6, 10.37, 0.001, and 0.34, respectively. The average reputation, tenure, and morning are 39491, 2210.4, and 0.41, respectively. They show that an answer, on average, comes from a user with an average reputation of 39491, existing 2210.4 days, and is generated in the morning with a 41% chance.

## 4. Estimation strategy

In this research, we are interested in investigating the relative impact of GAI vs. humans on voluntary knowledge contribution on Q&A platforms. We assess the impact using two types of outcome variables: (1) the number of upvotes received; (2) and the textual quality of the answer in terms of answer length, readability, sentiment, and subjectivity.

To accomplish this, we use fixed-effects regression models, treating whether the answers are generated by GAI as an independent variable and controlling for answer-level and answer-generator-level fixed effects. However, the use of GAI to generate an answer may not be exogenous, raising potential endogeneity issues. To address the potential selection to use GAI for a particular answer, we conduct propensity score matching (PSM) and coarse exact matching (CEM) to mitigate the potential biases, following Khurana et al. (2019) and Ge et al. (2021).

## 5. Empirical results

In this section, we present our empirical results, including the main results, endogeneity checks, and results related to the exploration of the moderating mechanism of user reputation.

## 5.1. Main results

To compare the relative impact of GAI vs. humans on voluntary answer generation on knowledge contribution platforms, we operationalize the impact on two dimensions: answer community acceptance and text quality. Our main dependent variables are answer score, length, readability, sentiment, and subjectivity. The dependent variables are measured at the answer level. Following the literature (e.g., Wu 2013), we run a fixed effects model to empirically examine the main effects. The model specifications are shown in the following equation (1).

$$Y_i = \beta_0 + \beta_1\, GAI_i + \beta_2\, Controls + u_j + q_k + \varepsilon_{ijk} \quad (1)$$

where $Y_i$ indicates a dependent variable of the answer $i$; $GAI_i$ is a dummy variable, representing whether the answer is generated by the GAI (= 1, Yes; = 0, No); $Controls$ mainly reveals the user-question variant control variables, which include $Tenure_{ij}$ and $Morning_{ij}$; $u_j$ and $q_k$ represent the user and question-level fixed effects, respectively. The definitions of the variables used are stated in Table 1. The estimation results are presented in Table 3.

**Table 3. Estimation Results of the Main Model**

| | (1) score | (2) Length | (3) GF | (4) Sentiment | (5) Subjectivity |
|---|---|---|---|---|---|
| GAI | -0.02** | -7.30*** | -0.78*** | 0.003+ | -0.03*** |
| | (0.01) | (0.31) | (0.12) | (0.00) | (0.00) |
| Tenure | -0.08*** | 0.18 | 0.02 | 0.01** | 0.002 |
| | (0.01) | (0.46) | (0.15) | (0.00) | (0.01) |
| Morning | 0.03* | 0.35 | -0.31+ | -0.004 | -0.0034 |
| | (0.01) | (0.58) | (0.18) | (0.00) | (0.01) |
| GF | | 0.44*** | | | |
| | | (0.07) | | | |
| Length | | | 0.07*** | | |
| | | | (0.00) | | |
| Constant | 165.3*** | -390.2 | -38.25 | -16.15** | -5.45 |
| | (22.09) | (1006.74) | (327.24) | (5.23) | (13.73) |
| UFE | Yes | Yes | Yes | Yes | Yes |
| PFE | Yes | Yes | Yes | Yes | Yes |
| Obs | 40023 | 40023 | 40023 | 40023 | 40023 |
| R | 0.33 | 0.17 | 0.20 | 0.12 | 0.14 |

***Note.*** *User cluster standard errors in parentheses; + p < 0.1, * p<0.05, ** p<0.01, *** p<0.001; the GF means the years of formal education needed to understand the generated answers; UFE = user fixed effects, PFE = post fixed effects, Obs = observation, and R = adjusted R-square; we include GF as a control for length and length as a control for GF because they are highly correlated.*

These empirical results address the interests of RQ1 and H1. Specifically, GAI contributes differently to answer quality than human users. When using GAI tools, answers, on average, receive lower upvote scores, are shorter in length, are easier to read, and are more positive and objective. One possible mechanism behind these results is that GAI, such as ChatGPT, occasionally produces incorrect answers and fabricated facts (Borji 2023), which may cause their solutions to receive fewer upvotes on average on SO platforms compared to human solutions. In addition, users' motivations for voluntarily contributing answers on Q&A platforms include reciprocity and gaining a social reputation (Chen et al. 2018). Therefore, human users may tend to generate more detailed answers with higher quality because they want to gain a social reputation from their contribution. Furthermore, when human users write answers to questions on SO, they may tend to provide suggestions, solutions, and recommendations, which may be more critical and negative, as suggested by Shan and Rivera (2022). Thus, compared to human users, answers from GAI are, on average, shorter, easier to read, and more positive. In addition, GAI, such as ChatGPT, has been found to generate more objective and neutral content in many other domains, such as finance, medicine, and Wikipedia, while humans tend to use more subjective expressions (Guo et al. 2023). Similarly, the GAI on the SO platform generates more objective answers than human users.

## 5.2. Endogeneity and robustness checks

**5.2.1. PSM**. Self-selection to use GAI tools (e.g., ChatGPT) may bias the results. The decision to use GAI tools may depend on observable characteristics, such as users' tenure in answering questions, users' community reputation and upvotes received, whether the question contains monetary incentives (BountyYes: = 1, yes; = 0, no), and whether users answer questions in the morning. To minimize the effect of these observable characteristics, we used propensity score matching (PSM) to re-align the responses coming from human users and the GAI. Following the literature, we matched the control group (responses coming from human users) to the treatment group (responses coming from GAI tools) using a one-to-one, nearest-neighbor selection process. For each unit in the treatment group, we found the best-matching response from the control group in terms of observed characteristics. After matching, the usable sample size was reduced from 40,023 to 33,430.

The PSM performance and statistics reported in Figure 2(a) show that the imbalance of all covariates was significantly reduced after matching, establishing that the treatment and control groups were well-matched in terms of observable characteristics. Figure 2(b) shows the propensity score histogram, noting the overlapping distributions between the two groups. We reran the regression analysis on this modified sample, estimating equation (1). The results reported in Table 4 show that all of our findings are robust, i.e., answers from the GAI tools are associated with fewer upvotes, shorter length, easier to read, more positive, and less subjective content.

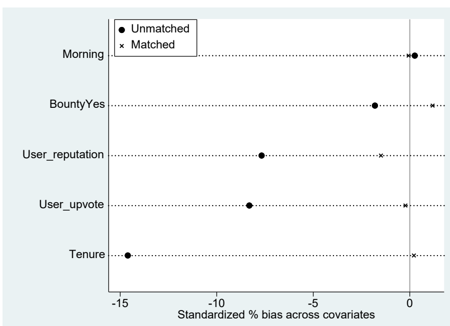


**Figure 2. Standardized Percentage Bias.**

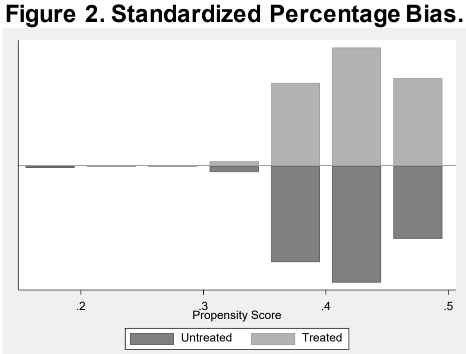


**Figure 3. Propensity Score Histogram.**

**Table 4. Estimation Results of the PSM Model**

| | (1) score | (2) Length | (3) GF | (4) Sentiment | (5) Subjectivity |
|---|---|---|---|---|---|
| GAI | -0.02** | -7.32*** | -0.78*** | 0.004** | -0.02*** |
| | (0.01) | (0.24) | (0.09) | (0.00) | (0.00) |
| Tenure | -0.06*** | -0.42 | -0.24 | 0.01* | -0.01 |
| | (0.01) | (0.39) | (0.15) | (0.00) | (0.01) |
| Morning | 0.03** | 0.76 | -0.54** | -0.002 | -0.001 |
| | (0.01) | (0.54) | (0.21) | (0.00) | (0.01) |
| GF | | 0.51*** | | | |
| | | (0.02) | | | |
| Length | | | 0.07*** | | |
| | | | (0.00) | | |
| Constant | 123.2*** | 887.6 | 500.2 | -10.98* | 17.48 |
| | (19.10) | (797.61) | (304.59) | (4.52) | (12.67) |
| UFE | Yes | Yes | Yes | Yes | Yes |
| PFE | Yes | Yes | Yes | Yes | Yes |
| Obs | 33430 | 33430 | 33430 | 33430 | 33430 |
| R | 0.64 | 0.58 | 0.56 | 0.45 | 0.46 |

***Note.*** *User cluster standard errors in parentheses; + p < 0.1, * p<0.05, ** p<0.01, *** p<0.001; the GF means the years of formal education needed to understand the generated answers; UFE = user fixed effects, PFE = post fixed effects, Obs = observation, and R = adjusted R-square; we include GF as a control for length and length as a control for GF because they are highly correlated.*

**5.2.2. CEM**. Following the same reasons for using PSM, we also conduct CEM to mitigate the self-selection of a user to use generative AI tools to answer a question based on potential observed confounders. Compared to PSM, CEM can produce matched samples with lower covariate imbalance (Ge et al. 2021). Specifically, CEM creates meaningful bins for each covariate, matches the treatment and control samples based on the bins, and keeps the original values of the covariates for analysis (Blackwell et al. 2019). Adopting this procedure by following Ge et al. (2021), we obtained 39,992 observations.

We reran the regression analysis across this modified sample, estimating equation (1). The results reported in Table 5 show that all our findings remain robust, i.e., answers from generative AI tools are associated with a decreased answer community upvote score, length, and subjectivity, but an easier-to-read and more positive content.

**Table 5. Estimation Results of the CEM Model**

| | (1) score | (2) Length | (3) GF | (4) Sentiment | (5) Subjectivity |
|---|---|---|---|---|---|
| GAI | -0.03*** | -8.00*** | -0.84*** | 0.003* | -0.04*** |
| | (0.01) | (0.22) | (0.09) | (0.00) | (0.00) |
| Tenure | 0.00*** | 0.001*** | 0.00*** | -0.00*** | 0.00 |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| Morni | 0.01 | -0.22 | 0.01 | -0.00 | -0.01 |

| | | | | | |
|---|---|---|---|---|---|
| ng | (0.01) | (0.40) | (0.13) | (0.00) | (0.01) |
| GF | | 0.46*** | | | |
| | | (0.05) | | | |
| Length | | | 0.08*** | | |
| | | | (0.00) | | |
| Constant | -0.21*** | 22.03*** | 3.207* | 0.10+ | 0.68*** |
| | (0.06) | (6.65) | (1.38) | (0.05) | (0.02) |
| UFE | Yes | Yes | Yes | Yes | Yes |
| PFE | Yes | Yes | Yes | Yes | Yes |
| Obs | 39992 | 39992 | 39992 | 39992 | 39992 |
| R | 0.17 | 0.23 | 0.10 | 0.05 | 0.08 |

***Note.*** *User cluster standard errors in parentheses; + p < 0.1, * p<0.05, ** p<0.01, *** p<0.001; the GF means the years of formal education needed to understand the generated answers; UFE = user fixed effects, PFE = post fixed effects, Obs = observation, and R = adjusted R-square; we include GF as a control for length and length as a control for GF because they are highly correlated.*

## 5.3. Moderation mechanism of users' community reputation

In order to understand the quality differences between voluntary answers from generative AI and human users, we explore the moderating mechanism from users' community reputation. Users' community reputation affects their voluntary contribution quality on Q&A platforms because one of users' voluntary contribution motivations is to gain peer recognition and social reputation (Chen et al. 2018). Hence, the community reputation may moderate the relative impact of generative AI and human users by moderating users' voluntary contribution. Therefore, to examine how users' community reputation moderates the relative impact of generative AI vs. human users in terms of answer quality on Q&A platforms, we utilized the following regression model to unpack the mechanism:

$$Y_i = \beta_0 + \beta_1\, GAI_i \times Reputation\_std_j + \beta_2\, GAI_i + \beta_3\, Controls + u_j + q_k + \varepsilon_{ijk} \quad (2)$$

The meanings of the variables are the same as in equation (1). $Reputation\_std_j$ is the standardized form of, $Reputation_j$, the community reputation of user $j$. $Reputation_j$ is measured by summarizing the increased reputation from a good contribution (e.g., a contributed answer is accepted or gains a bounty) and decreased reputation from a bad contribution (e.g., low-quality answer getting downvotes) (Cooper 2020). In our moderation, we use the standardized reputation because the raw reputation has a very large and diverse distribution (See Table 2). We also use the interaction term ($GAI_i \times Reputation_j$) to investigate how users' community reputation moderates the relative impact of generative AI tools vs. human users on the knowledge contribution in terms of answer generation quality, capturing that interaction effect with the estimation of $\beta_1$. We used control variables, including *Tenure*$_{ij}$ and *Morning*$_{ijk}$. The results are shown in Table 6.

**Table 6. Estimation Results of Interaction Effect of Users' Community Reputation.**

| | (1) score | (2) Length | (3) GF | (4) Sentiment | (5) Subjectivity |
|---|---|---|---|---|---|
| GAI × Reputation_std | -0.02* | -0.62* | 0.16 | -0.06+ | 0.002** |
| | (0.01) | (0.25) | (0.11) | (0.03) | (0.00) |
| GAI | -0.02** | -7.28*** | -0.79*** | -0.60*** | 0.00+ |
| | (0.01) | (0.31) | (0.12) | (0.03) | (0.00) |
| Tenure | -0.07*** | 0.19 | 0.02 | 0.00 | 0.01** |
| | (0.01) | (0.46) | (0.15) | (0.05) | (0.00) |
| Morning | 0.03+ | 0.34 | -0.31+ | -0.06 | -0.00 |
| | (0.01) | (0.58) | (0.18) | (0.07) | (0.00) |
| GF | | 0.44*** | | | |
| | | (0.07) | | | |
| Length | | | 0.07*** | | |
| | | | (0.00) | | |
| Constant | 164.7*** | -406.2 | -33.99 | 1.10 | -16.08** |
| | (22.08) | (1006.98) | (327.42) | (110.43) | (5.23) |
| UFE | Yes | Yes | Yes | Yes | Yes |
| PFE | Yes | Yes | Yes | Yes | Yes |
| Obs | 40023 | 40023 | 40023 | 40023 | 40023 |
| R | 0.33 | 0.17 | 0.20 | 0.15 | 0.12 |

***Note.*** *User cluster standard errors in parentheses; + p < 0.1, * p<0.05, ** p<0.01, *** p<0.001; the GF means the years of formal education needed to understand the generated answers; UFE = user fixed effects, PFE = post fixed effects, Obs = observation, and R =*

*adjusted R-square; we include GF as a control for length and length as a control for GF because they are highly correlated.*

Table 6 reports the results of using the users' community reputation as the moderation effect for the relative impact of generative AI vs. humans on voluntary knowledge contribution. We found that users' reputation moderates the relative impact. Specifically, by increasing the standard deviation of the community reputation, the relative impact of generative AI and humans is more pronounced for answer score and length while less salient for sentiment and subjectivity. Specifically, if human users have one standard deviation higher reputation, the relative impact of generative AI vs. human on answer score and length is enhanced by 0.0223 and 0.0559, respectively. Therefore, a higher level of community reputation, representing users gain more recognition and reputation from their prior contribution, can make their answers receive more upvotes from the community and have longer content. Interestingly, we find that the relative impact of the generative AI and human users on answer sentiment and subjectivity is weakened if users have a higher reputation. Specifically, if users have one standard deviation higher reputation, the relative impact of generative AI vs. human on answer sentiment and subjectivity is weakened by 0.0559 and 0.00274, respectively. That means users with higher community reputations tend to generate more positive and less subjectivity answers, making the sentiment and subjectivity difference between answers from generative AI and human users less salient. However, in terms of the relative impact on the generated answer readability, users' community reputation does not play a moderation role. Hence, H2 is partially supported.

These empirical findings address RQ2's and H2's interest in the moderating role of user reputation on the relative impact of generative AI vs. humans on users' voluntary knowledge contribution on Q&A platforms. Users having longer reputation indicates that theyhave more experience in contributing high-quality content, which can be easier to receive upvotes and recognition from other community users on the platforms (Chen et al. 2018) and form a regular habit of contributing in a high-quality manner (Kim et al. 2005). Additionally, a higher level of reputation can enhance users' motivation to gain recognition in contributing voluntary knowledge, which improves their contribution in a high-quality manner (Deodhar and Gupta 2023). Hence, for the users having a high community reputation, the quality differences of answers from generative AI and human users are expected to be much more different. Our empirical results supported that higher user reputation makes the answer score and length further different but make the sentiment and subjectivity less different. Interestingly, we did not find that the higher reputation moderates the relative impact of generative AI and human users on the answer readability. That means human users with higher reputations may not change the way (e.g., easier to read or hard to read) they write the answers.

## 6. Discussion

Voluntary knowledge in Q&A platforms is crucial for users, platforms, and organizations. With recent advances in generative artificial intelligence (AI) techniques, automatically generated knowledge has become possible. The development of GAI chatbots such as ChatGPT, which are based on large language models (e.g., GPT-3.5), has made it possible to automatically generate answers on Q&A platforms. However, it is unclear whether the answers generated by GAI are different from those generated by human users, especially how they differ for different users. To find out, this study examines the relative impact of GAI and human users on voluntary knowledge contribution in terms of answer quality measures: score, length, readability, sentiment, and subjectivity. We also examine how users' community reputation moderates the relative impact of GAI and human users.

Using a dataset from one of the largest Q&A platforms and employing fixed-effects models with both propensity score matching and coarse exact matching econometric techniques to mitigate potential endogeneity issues, we find that, relative to human users, the GAI produces answers with lower answer scores (i.e., receiving fewer upvotes), shorter, easier to read, more positive, and less subjective content. In addition, we find that users with higher community reputation increase the relative impact of the GAI and human users on answer score and length but weaken the relative impact on sentiment and subjectivity, i.e., users with higher community reputation produce answers receiving more upvotes, longer content, more positive, and less subjective tone. However, the effect on the readability of the answer does not change. This means that users with higher community reputations do not change content generation in terms of readability.

### 6.1. Theoretical and practical contribution

Our research enriches several streams of IS literature. First, we enrich the IS literature on voluntary knowledge contribution. Specifically, we are among the first to investigate the relative influence of GAI and human users on voluntary knowledge contribution on knowledge platforms. We aim to understand whether GAI produces answers differently than human users and how user characteristics such as reputation moderate the relative impact. Second, we enrich the IS literature on the relative impact of AI and humans. In particular, we enrich the understanding of the relative impact of GAI vs. humans in the context of answering a question on Q&A platforms.

In addition to contributing to the literature, our research also provides several managerial implications. Our research findings are important for knowledge platform owners and managers to consider as they develop strategies for managing users' use of GAI on their platforms to encourage high-quality knowledge contributions. In addition, our research findings inform users and organizations regarding the quality differences in voluntary knowledge contributions between generative AI and human users. Specifically, platform owners and managers, users, and organizations are informed that content produced by GAI will receive lower response scores, as well as shorter, more positive, and less subjective responses compared to an average human user. Also, the relative response differences between GAI and human users will be different for different users. For example, if the users have a high community reputation, the relative differences in answer score and length will be more pronounced, while answer sentiment and subjectivity will be less pronounced, and answer readability will not change. Being aware of these findings and the dynamics of the differences between GAI and human users in voluntary knowledge contribution, platform managers and owners can develop better strategies for managing the use of GAI on their platforms, which is critical because knowledge platforms such as Stack Overflow has banned the usage of GAI on their platform. Furthermore, being aware of these findings, users can make a better decision of using GAI or answering themselves to gain more benefits such as gaining more job opportunities or reputation, and organizations can make a better decision in terms of encouraging their employees to generate answers themselves or using GAI to gain more benefits such as improving the productivity.

### 6.2. Limitations and future research

Our research aims to understand the relative influence of GAI and human users on users' voluntary contribution of answers on knowledge platforms. This research is not without its limitations. First, we focus only on the Stack Overflow platform, which may limit the generalizability of the research findings. Future research can utilize data from other knowledge contribution platforms.

Second, when operationalizing the use of GAI, we only used generative answer generation detection based on the GPT-2 output detector developed by OpenAI. Although this method achieves high accuracy (e.g., 99.3% accuracy), there may be some misidentifications of the answers using GAI. Therefore, one of our future research studies is to design online experiments to manipulate whether the answer was generated by a GAI or a human user.

Finally, regarding the moderation mechanism, we only looked at user reputation. There may be other potential moderators from a different perspective, such as user tenure and monetary bounty. Our future research will further explore these moderators to further unpack the mechanisms of the relative influence of GAI and human users.

## 7. References


Abdullah, U. (2023). How accurate is Chat GPT in generating responses? *PC Guide*. (March 7). https://www.pcguide.com/apps/how-accurate-is-chat-gpt/.

Alcántara, A. (2023). Is it human or AI? New tools help you spot the bots. *The Wall Street Journal*, (January 10). https://www.wsj.com/articles/is-it-human-or-ai-new-tools-help-you-spot-the-bots-11673356404.

Bateman, P. J., Gray, P. H., & Butler, B. S. (2011). Research note—the impact of community commitment on participation in online communities. *Information Systems Research*, 22(4), 841-854.

Blackwell, M., Iacus, S., King, G., & Porro, G. (2009). CEM: Coarsened exact matching in Stata. *The Stata Journal*, 9(4), 524-546.

Borji, A. (2023). A categorical archive of ChatGPT failures. *arXiv preprint arXiv:2302.03494*.

Bowman, E. (2023). A college student created an app that can tell whether AI wrote an essay. *NPR*, (January 9). https://www.npr.org/2023/01/09/1147549845/gptzero-ai-chatgpt-edward-tian-plagiarism.

Burtch, G., He, Q., Hong, Y., & Lee, D. (2022). How do peer awards motivate creative content? Experimental evidence from Reddit. *Management Science*, 68(5), 3488-3506.

Chen, W., Wei, X., Zhu, K. X. (2018). Engaging voluntary contributions in online communities: A hidden Markov model. *MIS Quarterly*, 42(1), 83-100.

David, C. (2023). Stack Overflow growth and usage statistics (2023). *SignHouse*, (January 17). https://www.usesignhouse.com/blog/stack-overflow-stats.

Deodhar, S. J., & Gupta, S. (2023). The impact of social reputation features in innovation tournaments: Evidence from a natural experiment. *Information Systems Research*, 34(1), 178-193.

Gates, B. (2023). https://www.gatesnotes.com/The-Age-of-AI-Has-Begun.

Ge, R., Zheng, Z., Tian, X., & Liao, L. (2021). Human–robot interaction: When investors adjust the usage of robo-

advisors in peer-to-peer lending. *Information Systems Research*, 32(3), 774-785.
Guo, X., Shan, G., Rivera, M., & Qiu, L. (2022). Improving response to digital nudges: Social norm appeals in real-time feedback applications. *Available at SSRN 4097025*.
Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.
Huang, P., Ceccagnoli, M., Forman, C., & Wu, D. J. (2022). IT knowledge spillovers, absorptive capacity, and productivity: Evidence from enterprise software. *Information Systems Research*, 33(3), 908-934.
Jia, N., Luo, X., Fang, Z., & Liao, C. (2023). When and how artificial intelligence augments employee creativity. *Academy of Management Journal*. Forthcoming.
Jin, J., Zhang, T., & Yan, X. (2023). Why do users continually seek knowledge in online Q&A communities? An empirical investigation. *Information Discovery and Delivery*, 51(1), 1-12.
Khern-am-nuai, W., Kannan, K., & Ghasemkhani, H. (2018). Extrinsic versus intrinsic rewards for contributing reviews in an online platform. *Information Systems Research*, 29(4), 871-892.
Khurana, S., Qiu, L., & Kumar, S. (2019). When a doctor knows, it shows: An empirical analysis of doctors' responses in a Q&A forum of an online healthcare portal. *Information Systems Research*, 30(3), 872-891.
Kim, S. S., Malhotra, N. K., & Narasimhan, S. (2005). Research note—Two competing perspectives on automatic use: A theoretical and empirical comparison. *Information Systems Research*, 16(4), 418-432.
Kirchner, J. H. (2023). New AI classifier for indicating AI-written text. *OpenAI*, (January 31). https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.
Lee, S. Y., Rui, H., & Whinston, A. B. (2019). Is best answer really the best answer? The politeness bias. *MIS Quarterly*, 43(2), 579-600.
Luo, X., Qin, M. S., Fang, Z., & Qu, Z. (2021). Artificial intelligence coaches for sales agents: Caveats and solutions. *Journal of Marketing*, 85(2), 14-32.
Malgonde, O. S., Saldanha, T. J., & Mithas, S. (2023). Resilience in the open source software community: How pandemic and unemployment shocks influence contributions to others' and one's own projects. *MIS Quarterly*, 47(1), 361-390.
Mousavi, R., Raghu, T. S., & Frey, K. (2020). Harnessing artificial intelligence to improve the quality of answers in online question-answering health forums. *Journal of Management Information Systems*, 37(4), 1073-1098.
OpenAI. (2022). Introducing ChatGPT. *OpenAI*, (November 30). https://openai.com/blog/chatgpt.
Qiao, D., Lee, S. Y., Whinston, A. B., & Wei, Q. (2020). Financial incentives dampen altruism in online prosocial contributions: A study of online reviews. *Information Systems Research*, 31(4), 1361-1375.
Cooper, R. (2020). How to get your first 1000 reputation on Stack Overflow. https://robertcooper.me/post/1000-stack-overflow-reputation.
Shan, G., Zhang, D., Zhou, L., Suo, L., Lim, J., & Shi, C. (2018). Inconsistency investigation between online review content and ratings. In *Twenty-fourth Americas Conference on Information Systems*.
Shan, G., Zhou, L., & Zhang, D. (2021). From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. *Decision Support Systems*, 144, 113513.
Shan, G., & Rivera, M. (2022). Humanizing digital transformation across people and things: An empirical investigation of the impact of real-time feedback types on performance. *International Conference on Information Systems*.
Shan, G., Wattal, S., & Thatcher, J. B. (2022). How does anonymizing crowdsourced users' identity affect fact-checking on social media platforms? A regression discontinuity analysis. *International Conference on Information Systems*.
Shan, G., & Qiu, L. (2023). Examining the impact of generative AI on users' voluntary knowledge contribution: Evidence from a natural experiment on Stack Overflow. *Available at SSRN 4462976*.
Shan, G., & Jia, S. (2024). Generative AI or real users? Investigating the relative impact of generative AI vs. humans on online review quality. *Hawaii International Conference on System Sciences*.
Shearer, A. (2022). I used ChatGPT to answer questions on Stack Overflow. *DEV*, (December 4). https://dev.to/andrews1022/ive-been-using-chatgpt-to-answer-questions-on-stack-overflow-2ib3.
Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (2019). Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.
Stack Overflow. (2022). Temporary policy: ChatGPT is banned. *Stack Overflow*, (December 5). https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned.
Wang, J., Li, G., & Hui, K. L. (2022a). Monetary incentives and knowledge spillover: Evidence from a natural experiment. *Management Science*, 68(5), 3549-3572.
Wang, M., Shan, G., & Thatcher, J. B. (2022b). Human versus AI? Investigating the heterogeneous effects of live streaming E-commerce. *Americas Conference on Information Systems.*
White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*.
Wu, L. (2013). Social network effects on productivity and job security: Evidence from the adoption of a social networking tool. *Information Systems Research*, 24(1), 30-51.
Xu, L., Nian, T., & Cabral, L. (2020). What makes geeks tick? A study of Stack Overflow careers. *Management Science*, 66(2), 587-604.