Generative AI or Real Users? Investigating the Relative Impact of Generative AI vs. Humans on Online Review Quality

Guohou Shan
Temple University
guohou.shan@temple.edu

Shizhen Jia
University of Florida
shizhen.jia@ufl.edu

Abstract

Online reviews matter for customers, firms, and platforms increasingly. The recent advancement of generative Artificial Intelligence (AI) techniques makes it possible to generate online reviews automatically. However, the relative impact of generative AI vs. humans on online review generation is unknown. On the one hand, generative AI can generate high quality reviews because they are trained on diverse and high-quality data. On the other hand, generative AI hallucinates and may generate fabricated content, threatening the quality of the generated reviews. Using data from one of the biggest online review platforms, Yelp.com, we apply fixed effect models to understand the relative impact of generative AI vs. humans on the quality of generated reviews. We find that reviews from generative AI averagely have bigger ratings, a higher level of inconsistency between rating and sentiment, shorter, harder to read, and more positive and subjective content. Our study has both theoretical and practical implications.

Keywords: Online reviews, generative AI vs. humans, rating, review quality, fixed effects.

1. Introduction

Online consumer reviews (OCRs) have become increasingly important for customers, firms, and platforms. Customers rely on OCRs to learn more about their interested products and assist their purchasing decisions (Nie et al., 2022). OCRs have been found to impact firms' reputation and revenue to a large degree (Yang et al., 2023). For online review platforms, such as Yelp.com, OCRs play a crucial role in maintaining users' participation, which is often a challenge for online communities (Chen et al., 2018).

In recent years, significant progress in generative AI techniques, such as ChatGPT, has enabled the generation of online reviews through automation. ChatGPT, built on LLM like GPT-3.5 and incorporates reinforcement learning from human feedback, is capable of mimicking human behavior and producing comments for stores on online platforms in a conversational manner (OpenAI, 2023).

However, the relative impact of generative AI vs. human users on the online review generation has yet to be explored. On the one hand, generative AI has the potential to generate high quality reviews. Models such as ChatGPT are trained on diverse and high-quality data and are consistently refined based on human feedback (Pocock, 2023). These models have been shown to provide detailed responses (Guo et al., 2023).

On the other hand, generative AI has been confirmed as having hallucinations and may generate fabricated content (Susarla et al., 2023). That generated content is of low quality, which can harm the online review platform ecosystems. Understanding the relative impact of generative AI and humans on the generated review quality is crucial because it has implications for users, platforms, and firms. Therefore, we aim to address the following research question.

RQ1: *Does generative AI produce online reviews with different quality from real human reviews?*

The answer to this question is highly relevant to the industry practice. By addressing RQ1, we provide guidance for online review platform managers and owners in terms of using generative AI on their platforms to enhance platform development.

Our study sheds light on RQ1 by revealing that the OCRs generated from generative AI and human users are different in terms of various dimensions. Specifically, OCRs from generative AI tend to have bigger ratings, a higher level of inconsistency between star rating and review sentiment, and shorter, harder to read, more positive, and more subjective content.

Additionally, it is currently unclear whether OCRs produced by generative AI may be further different from OCRs from humans for stores with different characteristics, such as the overall store rating.

Previous research has shown that stores with higher overall ratings are more likely to have a loyal customer base (Anderson and Srinivasan, 2003). As a result, stores with higher overall ratings are more likely to have a higher proportion of positive reviews. Since the generative AI is trained on large datasets, the corpus related to stores with higher overall ratings may tend to be similarly positive, making the generated OCRs more similar to human generated OCRs.

URI: https://hdl.handle.net/10125/106873
978-0-9981331-7-1

Also, previous research (e.g., Tao and Kim, 2022; Weathers et al., 2015) finds that product types can moderate OCR generation and credibility. Since different types of stores may provide various products, the store types can moderate the OCR generation. Besides, humans and AI are perceived as having different levels of expertise under different contexts (Longoni and Cian, 2022). For example, AI is better at giving recommendations in a utilitarian context, while humans are better at giving suggestions in a hedonic and experiencing context. Hence, it is anticipated that generative AI may give different reviews compared with human users for different stores, which can provide utilitarian or experienced products or services. Therefore, to understand how stores' characteristics may moderate the relative impact of generative AI and humans on the quality of generated reviews, we propose our second research question as follows:

RQ2: *how do store characteristics moderate the relative impact of generative AI and humans on the quality of generated reviews?*

Answering RQ2, we provide online review platforms' managerial insights on when to adopt generative AI to increase the OCRs on their platform. Specifically, we find that overall store rating weakens the relative impact of generative AI and human users on the quality of generated reviews. Additionally, the relative impact of generative AI and human users differs for different stores.

This research makes a two-fold contribution. Theoretically, we contribute to the online review generation in Information Systems (IS) literature (e.g., Guan et al., 2023) and the relative impact of AI vs. human IS literature (e.g., Adam et al., 2022) by investigating the relative impact of generative AI and humans on the OCR generation on online review platforms. Practically, this research provides various managerial implications for online review platform managers to create policies that leverage generative AI on their platforms.

2. Literature review

2.1 Online customer reviews

OCRs refer to product evaluations created by customers and posted on company websites or third-party platforms such as (Mudambi and Schuff, 2010). These reviews serve as a valuable source of information for potential consumers, enabling them to make informed decisions and evaluate the quality and credibility of offerings (Chevalier and Mayzlin, 2006). They reduce information asymmetry between businesses and customers by providing a more balanced assessment of product quality and customer experience (Chevalier and Mayzlin, 2006). Several key concepts relate to OCRs, as illustrated in Table 1. Specifically, OCRs are generally composed of ratings, depth, valence, credibility, and disclosure information.

Table 1. The key concepts in OCRs

Concepts	Definitions
Ratings	A quantitative star rating (e.g., 1 to 5 stars). Ratings summarize the overall evaluation of a product or service (Mudambi and Schuff, 2010).
Review depth	The word count or length of a review. Deeper reviews tend to be more helpful to readers (Baek et al., 2012).
Review valence	The degree of positivity or negativity in a review. Positive reviews lead to higher product sales and prices (Liu, 2006).
Reviewer credibility	Credibility is signaled by the quality, depth, and consistency of reviews (Forman et al., 2008; Mudambi and Schuff, 2010)
Reviewer disclosure	Whether reviewers disclose their relationship to the business (Van Laer et al., 2013).

Table 2. Research topics of OCRs in IS research

Concepts	Empirical Discussions
Business performance	Positive online reviews increase sales and product prices. Negative reviews have a bigger impact on lesser-known brands (Chevalier and Mayzlin, 2006). Lu and Stepchenkova (2015) found that the number of reviews is positively associated with hotel revenue.
Review helpfulness	Review depth, the balance of positive and negative comments, and reviewer credibility affect review helpfulness (Baek et al., 2012). Disclosures enhance perceived helpfulness (Van Laer et al., 2013).
Reviewer motivations	Reviewers are motivated by both intrinsic reasons (altruism and enjoyment) and extrinsic reasons (social and economic rewards). Extrinsic motivations reduce review credibility (Hennig-Thurau et al., 2004).
Reviewer behaviors	Online customer reviews significantly affect consumers' purchase decision-making (Y. Chen and Xie, 2008).
Deception detection	Machine learning and language analysis techniques can detect deceptive opinions with reasonable accuracy (Huang, 2011). Deceptive reviews tend to be more extreme, emotional, and rhetorical (Huang, 2011).
Review manipulation	The problem of review manipulation, where businesses or individuals either post fraudulent reviews or remove negative reviews, has been another key area of research. Many empirical studies have shown that businesses can benefit significantly from posting fake positive reviews or deleting negative reviews (Mudambi and Schuff, 2010).

Over the years, the proliferation of online platforms has made it possible for customers to share their experiences, opinions, and evaluations of various products and services. As a result, OCRs have gained significant attention in IS research, including sales impact, helpfulness, motivation, and deception detection (as illustrated in Table 2). In terms of research methodology, there are three primary data analysis approaches: sentiment analysis (Liang et al., 2019; Picazo-Vela et al., 2010), opinion mining (Archak et al., 2011; Malik and Hussain, 2017), and text analytics (Ghose et al., 2012; Yin et al., 2014).

In summary, OCRs have gained much attention from IS research and become an essential source of information for consumers and businesses alike.

2.2 Generative artificial intelligence reviews

Generative artificial intelligence includes a class of

AI models focused on generating original content such as text, images, or other media. Although generative AI techniques have been around for years, the development of ChatGPT prompted a wave of conversations and debates across various fields (Susarla et al., 2023).

A generative AI review is a text-based comment on a product or service generated by an AI algorithm. The algorithm is trained on a large dataset of human-written reviews, and it uses this data to learn the patterns and features associated with positive and negative reviews. Once the algorithm has been trained, it can generate new reviews for products or services that humans have not yet reviewed (Cao et al., 2023). This research used ChatGPT as a fundamental AI algorithm to generate reviews. First, ChatGPT demonstrates a high level of language understanding and generation capabilities. Second, ChatGPT benefits from its extensive pre-trained knowledge, allowing it to generate reviews that incorporate a wide range of information and domain-specific knowledge. This pre-trained knowledge enhances the quality and relevance of the generated reviews. By acknowledging the associated risks (e.g., ethical issues) and implementing mitigation strategies, researchers can harness the capabilities of ChatGPT to generate high-quality reviews for their research, enabling deeper insights and analysis (Dwivedi et al., 2023).

2.3 Research gaps in current literature

While research on online reviews and their impact has been conducted (Baek et al., 2012; Luca and Zervas, 2016), several significant research gaps still need to be addressed. First, the existing literature has not explored the differences between online reviews generated by generative AI and those written by real human users. The existence of this research gap holds immense importance because it pertains to the credibility and trustworthiness of online reviews and AI tools, which greatly influence the decision-making of consumers. As for OCRs, they provide firsthand experiences and opinions of real customers, but they may also be subject to biases or manipulation. While generative AI tools, such as ChatGPT, offer scalability, efficiency, and potential customization. However, they lack direct experiences and may not capture the nuanced contextual understanding found in human-generated reviews.

Moreover, previous studies have discussed the overall impact of AI-generated reviews and the potential biases or ethical concerns associated with them (Dwivedi et al., 2023; Susarla et al., 2023). However, limited empirical research directly compares the quality of reviews generated by AI systems with those written by humans. Understanding these differences is essential for assessing the credibility and trustworthiness of AI-generated reviews.

Second, while some studies have examined the influence of various store characteristics on the impact of OCRs (Utz et al., 2012; Xu, 2020), there is a dearth of research investigating how these characteristics influence generative AI reviews, especially the comparison between generative AI and customer reviews. Store characteristics such as reputation or type may interact with the source of reviews (generative AI or humans) and influence consumer perceptions and behaviors differently.

By addressing these research gaps, researchers can better understand the unique characteristics of AI-generated reviews compared to those authored by humans. Moreover, this research provides insights into the contextual factors shaping consumer responses to online reviews, ultimately informing businesses and policymakers on optimizing online review systems.

3. Research model and hypotheses

In this section, we discuss our research model and put forward our hypotheses on the effects of the relative generative AI vs. humans on the generated online reviews.

3.1. Mechanisms of the relative impact of generative AI and OCRs

Previous research shows that online reviews significantly impact consumer perceptions and purchase decisions (Chevalier and Mayzlin, 2006; Liu, 2006). While most review platforms contain authentic customer opinions, some may manipulate reviews to mislead consumers for commercial gains (Luca and Zervas, 2016; Mayzlin et al., 2014). With recent advances in natural language generation (NLG) and generative adversarial networks (GANs), it is possible to automatically generate fake reviews that seem authentic (Kryściński et al., 2019). Such generated AI reviews have the potential to manipulate consumer perceptions if not properly identified and regulated. Since the goal of NLG models and GANs is to generate reviews that seem authentic, the AI may provide overly enthusiastic reviews and ratings to mask their authenticity, as found in some promotional human reviews (Mayzlin et al., 2014). This could lead to systematic inflation of ratings in AI reviews compared to real customer opinions. Therefore, we hypothesize that:

H1a: Generative AI reviews, on average, have higher ratings compared to human generated OCRs.

While ratings provide an overall summary evaluation, the review text details the reviewer's opinions and sentiments. For authentic reviews, there should be consistency between the rating and sentiments (Yin et al., 2014). However, AI tools may not be able to precisely capture such consistency due to limitations in natural language understanding. The sentiments expressed in the review text may sometimes contradict the chosen rating, signaling the non-human origin of the review. Thus, we propose the following hypothesis:

H1b: Generative AI reviews exhibit a higher level of inconsistency between the rating and the sentiment expressed compared to OCRs.

Previous research shows that textual quality significantly impacts online reviews' perceived credibility and helpfulness. Higher quality reviews tend to be more coherent and objective and contain richer details (Korfiatis et al., 2012; Mudambi and Schuff, 2010)). While AI has achieved human-level performance on many natural language generation tasks, its output may still lack some attributes of human writing that signal high quality. Studies found that machine-generated text can appear "robotic", repetitive, and impersonal compared to human writing (Hill et al., 2015). AI models may reuse phrases or produce generic, simplistic accounts that seem less authentic (Sheng et al., 2019). Even state-of-the-art GANs trained on large datasets generate text with statistically different patterns compared to human language (Bender et al., 2021). These differences suggest that AI-generated reviews may exhibit poorer textual quality relative to customer opinions.

Overall, while AI progress points to its increasing ability to generate persuasive reviews (Brown et al., 2020), limitations remain around replicating the richness, personal experiences, and objectivity in human writing. The lack of "human touch" may be detectable in AI reviews, signaling their machine origin. Hence, we propose the following:

H1c: Generative AI reviews, on average, have a lower level of textual quality compared to OCRs.

3.2. Moderating impact of stores' characteristics

Previous research has shown that businesses with higher overall ratings are more likely to have a loyal customer base (Anderson and Srinivasan, 2003). One possible reason is that loyal customers are more likely to be satisfied with the business and its products or services, and they are more likely to share their positive experiences with others. Since AI models rely on patterns in the data to generate new samples, when the data is predominantly positive, the models will likely learn these patterns and generate positive samples. However, when data is more mixed or polarized, accurately capturing the diversity of opinions and contextual details may prove more challenging for AI. As a result, businesses with higher overall ratings are more likely to facilitate generative AI to produce reviews more similar to human reviews. Accordingly, we propose the following hypothesis:

H2: The relative difference between AI-generated and OCRs will be lower for businesses with higher overall ratings.

Previous research has examined the moderating effect of different types of products. For example, Weathers et al. (2015) found that product type can moderate the relationship between review credibility and review helpfulness. Different product types require varying levels of domain expertise for accurate evaluation. For instance, technology products may necessitate detailed technical knowledge, while food items may require subjective assessments. Furthermore, Longoni and Cian (2022) confirmed that individuals are more likely to believe AI's recommendation on utilitarian product type rather than hedonic. Therefore, we hypothesize that:

H3: The type of store has a moderating effect on the relative impact between reviews from generative AI and OCRs across different review quality measures.

4. Research context and data

4.1. Research context

We use Yelp (https://www.yelp.com/) as our context to investigate our research questions and research model. Yelp has become nearly synonymous with online reviews since its inception in 2014. It has over 0.2 billion online reviews covering restaurants, shopping, food, etc. (Andre, 2023).

OpenAI launches the ChatGPT, a generative AI with large language models, on November 30, 2022. It is based on large language models and incorporates reinforcement learning from human feedback, capable of mimicking human behavior in generating content such as online reviews (OpenAI, 2023).

The launch of ChatGPT provides an opportunity to investigate the relative impact of generative AI and humans on the quality of produced online reviews. We ask ChatGPT to comment on particular stores as the online reviews in the treatment group with reviews from generative AI and reviews for the same stores from human customers as the control group. Having both treatment and control groups, we can empirically investigate the relative impact of generative AI and humans on online review generation quality.

4.2. Data

To investigate our research questions and model, we collect data from Yelp.com. Our target datasets come from the released loved brands in the U.S. (Yelp, 2023). Specifically, we select top-2 brands from each category (Food/Restaurants/Shopping). Then, we retrieve stores from those selected brands and their online reviews from the Yelp academic dataset (https://www.yelp.com/dataset) as the reviews from human users. Next, we utilize the ChatGPT API to generate the same number of online reviews with star ratings for a particular store associated with a selected brand in our dataset. When utilizing the ChatGPT API, we also set a high-temperature parameter of the ChatGPT to make the generated reviews as random as possible, allowing online reviews for a particular store to have variations, following (Gilardi et al., 2023).

Our dataset contains 11,169 Yelp reviews for 18 food stores, 60 restaurants, and 26 shopping stores. Using the addresses of those stores, we ask ChatGPT to generate the same number of reviews (Our question or prompt to ChatGPT is, "Please generate a comment for [store name] at location [store address]"). Hence, our final data sample consists of 22,338 reviews, with half of them from generative AI and the other half coming from real human customers.

In this research, we focus on investigating the relative impact of generative AI and humans on three dimensions of online reviews: review rating, inconsistency between review rating and the textual review sentiment, and textual review quality. We measure the inconsistency between the review rating and the textual review sentiment by following Shan et al. (2021). To measure textual review quality, we use the review length, readability, sentiment, and subjectivity, which are commonly used in prior research, to indicate the quality of online reviews (Shan et al., 2021; Zhang et al., 2016). We also explore moderating effects from the characteristics of the store. Thus, our data is composed of those variables shown in Table 5. To gain a better understanding of our data, we provide data summary statistics in Table 6 and conduct a correlation analysis of variables, where we found the review rating and readability (r = 0.57), business star and food (r = 0.60), food and restaurant (r = -0.67), and restaurant and shopping (r = -0.61) have high correlations with all of the other variables having negligible correlations.

Table 5. Variables and Definitions

Category	Variables	Definitions
Independent Variable	*$GAIH_i$*	It shows whether review *i* is generated by generative AI or human. = 1, generative AI; = 0, human.
Dependent Variables	*$Rating_i$*	It represents the star rating of review *i*.
	$Incon_i$	It represents the inconsistency measured by star rating and sentiment of review *i*. It is measured as the absolute difference between the z-scores of a rating and its sentiment. The higher the value is, the higher the inconsistency is.
	$Length_i$	It represents the length in terms of the number of words of review *i*.
	$Readability_i$	It represents the readability of review *i*. We used the gunning-fog measure, following Khern-am-nuai et al. (2018). It measures the years of formal education needed to understand the reviews.
	$Sentiment_i$	It represents the sentiment of review *i*. We use Python NLTK and lexicon, following Shan et al. (2018).
	$Subjectivity_i$	It represents the subjectivity of review *i*. We use Python TextBlob, following Mousavi et al. (2020).
Moderators	*SOR_j*	It represents the overall rating of store *j*.
	$Food_j$	It indicates whether store *j* belongs to the food category or not (=1, Yes; =0, No).
	$Restaurant_j$	It indicates whether store *j* belongs to the restaurant category or not (=1, Yes; =0, No).
	$Shopping_j$	It indicates whether store *j* belongs to the shopping category or not (=1, Yes; =0, No).

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

Table 6. Variable Summary Statistics

Variables	Obs	Mean	Std.Dev.	Min	Max
GAIH (1)	22338	.5	.5	0	1
Rating (2)	18525	4.04	1.12	1	5
Incon (3)	18525	.98	.72	0	10.08
Length (4)	22338	75.02	68.19	1	946
Readability (5)	22338	9.23	3.14	.4	84.97
Sentiment (6)	22338	.27	.28	-1.85	3.74
Subjectivity (7)	22338	.62	.12	0	1
SOR (8)	22338	3.85	.44	3	5
Food (9)	22338	.16	.36	0	1
Restaurant (10)	22338	.70	.46	0	1
Shopping (11)	22338	.14	.35	0	1

Note. Obs = the number of observations; the observations of Rating and Incon are less than 22338 because ChatGPT sometimes refuses to generate a star rating when producing reviews.

From Table 6, the mean value of GAIH is 0.5, indicating that 50% of observations (reviews) are generated by using generative AI. The mean rating, incon, length, readability, sentiment, and subjectivity of a review are 4.04, 0.98, 75.02, 9.23, 0.27, and 0.62, respectively. The average store overall rating, and the category of food, restaurant, and shopping are 3.85, 0.16, 0.70, and 0.14, respectively.

5. Empirical results

In this research, we are interested in examining the relative impact of generative AI vs. humans on the quality of generated online reviews. To achieve this, we employ fixed effect regression models by treating whether the review generated from generative AI or a human as an independent variable and control the store-level fixed effect.

5.1. Main results

The dependent variables are measured at a review level. To control for store-level fixed effects, we conduct a fixed effect model to examine the main impact empirically. The model specifications are shown in the following equation (1).

$$Y_i = \beta_0 + \beta_1\ GAIH_i + \beta_2\ Controls + s_j + \varepsilon_{ij} \qquad (1)$$

where Y_i indicates a dependent variable of the review i; $GAIH_i$ is a dummy variable, representing whether the review is generated by the generative AI (= 1, Yes; = 0, No); $Controls$ mainly reveals the store level variant control variables, which include readability for the measuring the impact on rating and incon, and rating for measuring the impact on the readability because rating and readability are highly correlated and we want to control the omitted variable bias by including the control variables; s_j represents the store-level fixed effect. The definitions of the variables used are stated in Table 5. The estimation results are presented in Table 7.

Table7. Estimation Results of the Main Model.

	(1) Rating	(2) Incon	(3) Length	(4) Readability	(5) Sentiment	(6) Subjectivity
GAIH	0.56**	0.28**	-46.80**	3.80**	0.10**	0.10**
	(0.06)	(0.02)	(1.72)	(0.10)	(0.01)	(0.00)
Readability	-0.024**	0.01**				
	(0.00)	(0.00)				
Rating				-0.14**		
				(0.03)		
Con	4.03**	0.79**	98.42**	7.89**	0.22**	0.57**
	(0.03)	(0.02)	(0.86)	(0.12)	(0.00)	(0.00)
SFE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	18525	18525	22338	18525	22338	22338
R	0.04	0.04	0.12	0.32	0.03	0.16

Note. *Store cluster standard errors in parentheses; ** p<0.05; Con= Constant, Obs = Observation, and R = adjusted R-square.*

Table 7 reveals that generative AI and human users generated different online reviews for the same stores. Specifically, compared with online reviews from human users, on average, reviews from generative AI have a 0.56 higher rating, 0.28 higher level of inconsistency between rating and sentiment, contain 46.80 fewer words, require 3.80 additional years of education to understand the reviews, have 0.10 more positive sentiment, and are more subjective by having 0.10 more subjectivity values. Hence, our research hypotheses H1a and H1b are supported and H1c partially supported, meaning that generative AI can produce online reviews with mixed quality. On the one hand, the review from generative AI can contain higher ratings and be more positive. However, the review from generative AI has a higher level of inconsistency between rating and sentiment, which signals low quality based on Shan et al. (2021), contains shorter content, is harder to read, and is more subjective.

These empirical findings address the interests of RQ1 and H1. When the reviews come from generative AI, they contain higher ratings and inconsistency between rating and sentiment, shorter, harder to read, more positive, and subjective. One possible mechanism behind those results is that generative AI is trained on a large corpus but does not have real fresh experience compared with human users. Hence, when writing online reviews, generative AI tends to give higher ratings with higher inconsistency between rating and sentiment to mask their authenticity, as found in some human promotional reviews (Mayzlin et al., 2014). And they can also write shorter content because of lacking experience in generating real comments for the stores. Furthermore, when human users write a review for a store, they may give their experiences and recommendations, and suggestions, which can be critical and more negative, as suggested by (Shan and Rivera, 2022). Thus, compared with human users, generative AI will produce more positive reviews. Interestingly, we also found that the reviews generated from generative AI can be harder to read and more subjective, meaning human reviews tend to write easier to understand reviews with more objective content. One possible explanation is that human users may have real experiences, making them able to use easier to understand language and be more objective in commenting on the stores.

5.2. Moderation Mechanisms

In order to understand the differences between reviews generated by AI and human users, we explore moderating mechanisms from the store characteristics, including both store overall star rating and store category. We introduce the results as follows.

5.2.1. Store overall star rating. A store's overall star rating indicates the overall quality of the store. Stores having high overall star ratings normally have high quality, whereas the users' reviews tend to be positive. However, it is empirically unknown whether it may affect the review generation process of generative AI. Therefore, to examine how the store's overall rating (SOR) moderates the relative impact of generative AI vs. human users in terms of the quality of generated reviews, we utilized the following regression model to unpack the mechanism:

$$Y_i = \beta_0 + \beta_1\ GAIH_i \times SOR_j + \beta_2\ \ GAIH_i + \beta_3\ Controls + s_j + \varepsilon_{ij} \quad (2)$$

The meanings of the variables are the same as in equation (1). SOR_j is the overall rating of store j. To capture the moderation effect, we use the interaction term ($GAIH_i \times SOR_j$) with the estimation of β_1. The results are shown in Table 8.

Table 8. Estimation Results of Interaction Effect of Store Overall Star Rating (SOR)

	(1) Rating	(2) Incon	(3) Length	(4) Readability	(5) Sentiment	(6) Subjectivity
GAIH × SOR	-0.81*	-0.17*	11.02**	-0.68**	-0.05**	-0.02**

	*	*				
	(0.03)	(0.06)	(3.94)	(0.15)	(0.01)	(0.01)
GAIH	3.69**	0.94**	-89.21**	6.43**	0.27**	0.18**
	(0.12)	(0.21)	(14.93)	(0.63)	(0.04)	(0.04)
Readability	-0.03**	0.01**				
	(0.00)	(0.00)				
Rating				-0.16**		
				(0.03)		
Con	4.05**	0.79**	98.42**	7.97**	0.22**	0.57**
	(0.03)	(0.02)	(0.76)	(0.12)	(0.00)	(0.00)
SFE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	18525	18525	22338	18525	22338	22338
R	0.07	0.05	0.12	0.32	0.03	0.16

Note. *Store cluster standard errors in parentheses; ** p<0.05; Con= Constant, Obs = Observation, and R = adjusted R-square.*

Table 8 reports the results of using the SOR as the moderation effect for the relative impact of generative AI vs. humans on the quality of generated reviews. We found that SOR moderates the relative impact. Specifically, by increasing a unit of the store's overall rating, the relative impact of generative AI vs. human on review rating, inconsistency between rating and sentiment, length, readability, sentiment, and subjectivity is reduced by 0.81, 0.17, 11.02, 0.68, 0.05, and 0.02. That means the differences in reviews between generative AI and human users are smaller for the stores having a higher overall rating. Hence, H2 is supported.

These empirical findings address RQ2's and H2's interest in the moderating role of the store's overall rating on the relative impact of generative AI vs. humans on the quality of generated reviews. Stores with higher star ratings indicate that stores provide high service quality. For those stores, reviews generated from generative AI and human users are less different. Our empirical results supported that stores with high overall ratings may have more loyal customers sharing more positive experiences (Anderson and Srinivasan, 2003). Due to the fact that generative AI models are trained on large public datasets (Cao et al., 2023), including reviews in different platforms for stores with high overall ratings, the generated reviews for those stores can be highly similar in terms of review rating, inconsistency between rating and sentiment, length, readability, sentiment, and subjectivity. Therefore, the differences between reviews from generative AI and humans are smaller for stores with high overall ratings.

5.2.2. Store category. Different stores, such as food, restaurant, and shopping, provide different types of services, which may make the user-generated reviews different in many dimensions, as suggested by Khern-am-nuai et al. (2018). To understand how it may moderate the relative impact of generative AI and human users in terms of generating reviews, we utilized the following model to unpack the mechanism:

$$Y_i = \beta_0 + \beta_1\, GAIH_i \times Food_j + \beta_2\, GAIH_i \times Shopping_j + \beta_3\, GAIH_i + \beta_4\, Controls + s_j + \varepsilon_{ij} \quad (3)$$

The meanings of the variables are the same as in equation (1). $Food_j$ represents whether the store j is a type of food store (=1, Yes; =0, No) and $Shopping_j$ represents whether store j is a type of shopping store (=1, Yes; =0, No), where we use restaurant stores as the baseline group. The interaction terms: $GAIH_i \times Food_j$ and $GAIH_i \times Shopping_j$ capture the moderation roles of food stores (compared with restaurants) and shopping stores (compared with restaurants) on the relative impact of generative AI and human users on the review generation. The moderation effects are captured by the estimation of β_1 and β_2. Table 9 shows the results.

Table 9. Estimation Results of Interaction Effect of Store Category

	(1) Rating	(2) Incon	(3) Length	(4) Readability	(5) Sentiment	(5) Subjectivity
GAIH × Food	-0.58**	-0.35**	-3.09	-0.56**	0.02*	0.01
	(0.07)	(0.03)	(3.28)	(0.15)	(0.01)	(0.01)
GAIH× Shopping	0.026	-0.10**	-21.71**	-0.44*	0.03*	-0.001
	(0.12)	(0.03)	(3.76)	(0.23)	(0.01)	(0.01)
GAIH	0.66**	0.36**	-43.19**	3.97**	0.09**	0.10**
	(0.07)	(0.02)	(1.99)	(0.13)	(0.01)	(0.00)
Readability	-0.025**	0.01**				
	(0.00)	(0.00)				
Rating				-0.15**		
				(0.03)		
Con	4.034**	0.80**	98.42**	7.92**	0.22**	0.57**
	(0.03)	(0.01)	(0.76)	(0.12)	(0.00)	(0.00)
SFE	Yes	Yes	Yes	Yes	Yes	Yes
Obs	18525	18525	22338	18525	22338	22338
R	0.05	0.05	0.12	0.32	0.03	0.16

Note. *Store cluster standard errors in parentheses; * p<0.1, ** p<0.05; Con= Constant, Obs = Observation, and R = adjusted R-square.*

From Table 9, we found that compared with restaurant stores, food stores make the relative impact of generative AI and human users less salient on the review rating, inconsistency, and readability but more pronounced on review sentiment. Furthermore, compared with restaurant stores, shopping stores make the relative impact of generative AI and human users more pronounced on the review length and sentiment while making the impact on the review inconsistency and readability less salient.

These empirical findings address RQ2's and H3's interest in the moderating role of store categories on the relative impact of generative AI vs. human users on the quality of generated online reviews. Food stores, restaurants, and shopping stores are three different types of stores serving different types of products and services. According to prior research (Tao and Kim, 2022), customer reviews are expected to be different for different products. Customers may mainly focus on food quality in the food stores (product) and shopping experience in the shopping stores (service), while they care about both food and experiences in restaurant stores (both product and service). Hence, customers may purposely write different reviews with a different focus for different stores. Also, prior research (Longoni and Cian, 2022) reveals that AI is better at giving recommendations in a utilitarian context, while humans are better at giving recommendations in a hedonic and experiencing context. Hence, it is anticipated that generative AI may give different reviews compared with human users for different stores.

6. Discussion

Online customer reviews are important for customers, firms, and platforms. However, it is unclear whether reviews produced by generative AI are different from those generated by human users, especially how they are different for different types of stores. To figure those out, this study examines the relative impact of generative AI and human users on the quality of generated reviews in terms of several dimensions: star rating, inconsistency between rating and sentiment, review length, readability, sentiment, and subjectivity. Also, we explore how stores' overall star ratings and types moderate the relative impact.

Using a dataset from one of the biggest online review platforms (i.e., Yelp.com) with augmentation of generated reviews from ChatGPT API and employing fixed effect models to remove some potential confounders, we find the generative AI produces OCRs containing bigger star ratings, higher level of inconsistency between rating and sentiment, shorter, harder to read, more positive, and more subjective content. Additionally, we find that a higher overall store rating weakens the relative impact of generative AI and human users. Also, the store types (i.e., food vs. restaurant vs. shopping) moderate the relative impact of generative AI and humans.

6.1. Theoretical and practical contribution

Our research enriches several streams of IS literature. First, our study serves to amplify the IS literature concerning online reviews, demonstrating our pioneering role in investigating the intricate interplay between generative AI and human users in shaping the quality of OCRs. This endeavor is of paramount importance as it seeks to unravel the nuanced distinctions in OCRs originating from generative AI systems versus human contributors. Concurrently, our focus extends to the moderating role that store characteristics, encompassing factors such as overall rating and categorization, might provide insights into theoretical frameworks of information asymmetry and online trust rebuilding. By illuminating these facets, our research actively enhances the IS literature's comprehension of the multifaceted landscape that shapes the ecosystem of online reviews, thereby deepening the understanding of their intricate dynamics. Second, our study significantly augments the IS literature's discourse on the interplay between AI and human actors. With a specialized emphasis on the burgeoning domain of generative AI, we delve into the realm of review generation within online platforms. This deliberate focus underscores our commitment to delving into the frontiers of technology-mediated human interactions and the novel role that generative AI plays in reshaping these interactions. By channeling our efforts into this unique context, we facilitate a comprehensive understanding of how generative AI integrates into the fabric of online platforms, subsequently enriching the broader discourse on AI-human dynamics within the IS literature.

Aside from contributing to the literature, our research also provides several managerial implications. Our research findings are important for online review platform owners and managers to consider when developing strategies for managing users' generative AI usage on their platforms in order to encourage the production of high-quality OCRs. Specifically, platform owners and managers are informed that reviews produced from generative AI may have mixed quality compared with human users. Also, the relative review differences from generative AI versus human users are different for different types of stores. With awareness of those findings and the dynamics of the differences between generative AI and human users in producing reviews, online review platform managers and owners can design better strategies to manage the usage of generative AI on their platforms.

6.2. Limitations and future research

Our research aims to understand the relative impact of generative AI and human users on the quality of generated online reviews. This research is not without limitations. First, we only focus on the Yelp platform, which may limit the generalizability of the research findings. Future research can leverage data from other online review platforms.

Second, we only explored the moderation of the store characteristics, such as the overall rating and the store type. There could be other potential moderators from another angle, such as human users' OCR generation tenure and or locations of stores. Our future research will further explore those moderators to furtherly unpack the mechanisms of the relative impact of generative AI and human users.

Third, this study only analyzes the impact of reviews generated by one specific version of an AI model, ChatGPT 3.5, released at a single point in time. However, generative language models are continually being upgraded and refined by their developers. Newer versions may demonstrate meaningfully different characteristics compared to earlier releases as the underlying technologies progress. Therefore, the findings here represent a snapshot in time and may not generalize to subsequent upgrades of ChatGPT or other generative AI that emerge. A dynamic research approach is needed to keep pace with technological advancement and assess how implications could change as capabilities evolve over time. While this initial examination provides insights into one model version's effects, future work must investigate multiple upgrades longitudinally to develop a more comprehensive understanding of generative AI's shifting impact on online reviews as technologies progress rapidly.

Last, our study's sample was derived from well-recognized and "loved brands" on Yelp. This deliberate selection may limit the generalizability of our findings to businesses with strong brand recognition. Reviews for such brands might be influenced by consumers' pre-existing familiarity with the products and services they offer, potentially affecting the perceived value of online reviews. Future research should extend the investigation to include businesses with varying levels of brand recognition, allowing for a comprehensive analysis of the interactions between online reviews and consumer perceptions across different brand contexts.

7. References

Adam, M., Roethke, K., & Benlian, A. (2022). Human versus automated sales agents: How and why customer responses shift across sales stages. *Information Systems Research*.

Anderson, R. E., & Srinivasan, S. S. (2003). E-satisfaction and e-loyalty: A contingency framework. *Psychology & Marketing*, *20*(2), 123–138.

Andre, L. (2023, April 26). *78 Yelp statistics you must know: 2023 market share & user profile analysis*. Financesonline.Com. https://financesonline.com/yelp-statistics/

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, *57*(8), 1485–1509.

Baek, H., Ahn, J., & Choi, Y. (2012). Helpfulness of online consumer reviews: Readers' objectives and review cues. *International Journal of Electronic Commerce*, *17*(2), 99–126.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). *A comprehensive survey of AI-generated content (AIGC): A history of Generative AI from GAN to ChatGPT* (arXiv:2303.04226). arXiv.

Chen, W., Wei, X., & Zhu, K. X. (2018). Engaging voluntary contributions in online communities: A Hidden Markov Model. *MIS Quarterly*, *42*(1), 83–100.

Chen, Y., & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science*, *54*(3), 477–491.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *XLIII*, 345–354.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, *19*(3), 291–313.

Ghose, A., Ipeirotis, P. G., & Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science*, *31*(3), 493–520.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). *ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks* (arXiv:2303.15056). arXiv.

Guan, Y., Tan, Y., Wei, Q., & Chen, G. (2023). When Images Backfire: The Effect of Customer-Generated Images on Product Rating Dynamics. *Information Systems Research*, isre.2023.1201.

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:2301.07597). arXiv.

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, *18*(1), 38–52.

Hill, J., Randolph Ford, W., & Farreras, I. G. (2015). Real

conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, *49*, 245–250.
Huang, S. (2011). Designing utility-based recommender systems for e-commerce: Evaluation of preference-elicitation methods. *Electronic Commerce Research and Applications*, *10*(4), 398–407.
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.
Korfiatis, N., García-Bariocanal, E., & Sánchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, *11*(3), 205–217.
Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019, August 23). *Neural Text Summarization: A Critical Evaluation*. ArXiv.Org.
Liang, X., Liu, P., & Wang, Z. (2019). Hotel Selection Utilizing Online Reviews: A Novel Decision Support Model Based on Sentiment Analysis and Dl-vikor Method. *Technological and Economic Development of Economy*, *25*(6), 1139–1161.
Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, *70*, 74–89.
Longoni, C., & Cian, L. (2022). Artificial Intelligence in Utilitarian vs. Hedonic Contexts: The "Word-of-Machine" Effect. *Journal of Marketing*, *86*(1), 91–108.
Lu, W., & Stepchenkova, S. (2015). User-generated content as a research mode in tourism and hospitality applications: Topics, methods, and software. *Journal of Hospitality Marketing & Management*, *24*(2), 119–154.
Luca, M., & Zervas, G. (2016). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science*, *62*(12), 3412–3427.
Malik, M. S. I., & Hussain, A. (2017). Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, *73*, 290–302.
Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, *104*(8), 2421–2455.
Mudambi & Schuff. (2010). Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly*, *34*(1), 185.
Nie, C., Zheng, Z. (Eric), & Sarkar, S. (2022). Competing with the sharing economy: Incumbents' reaction on review manipulation. *MIS Quarterly*, *46*(3), 1573–1602.
OpenAI. (2023). *Introducing ChatGPT*. https://openai.com/blog/chatgpt
Picazo-Vela, S., Chou, S. Y., Melcher, A. J., & Pearson, J. M. (2010). Why provide an online review? An extended theory of planned behavior and the role of Big-Five personality traits. *Computers in Human Behavior*, *26*(4), 685–696.
Pocock, K. (2023, May 10). *How accurate is Chat GPT In generating responses?* PC Guide. https://www.pcguide.com/apps/how-accurate-is-chat-gpt/
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, P. (2022). Humanizing Digital Transformation across People and Things: An Empirical Investigation of the Impact of Real-Time Feedback Types on Performance. *Forty-Third International Conference on Information Systems*, Copenhagen.
Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). *The Woman Worked as a Babysitter: On Biases in Language Generation* (arXiv:1909.01326). arXiv.
Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems. *Information Systems Research*.
Tao, S., & Kim, H.-S. (2022). Online customer reviews: Insights from the coffee shops industry and the moderating effect of business types. *Tourism Review*, *77*(5), 1349–1364.
Utz, S., Kerkhof, P., & van den Bos, J. (2012). Consumers rule: How consumer reviews influence perceived trustworthiness of online stores. *Electronic Commerce Research and Applications*, *11*(1), 49–58.
Van Laer, T., De Ruyter, K., & Cox, D. (2013). A Walk in Customers' Shoes: How Attentional Bias Modification Affects Ownership of Integrity-violating Social Media Posts. *Journal of Interactive Marketing*, *27*(1), 14–27.
Weathers, D., Swain, S. D., & Grover, V. (2015). Can online product reviews be more helpful? Examining characteristics of information content by product type. *Decision Support Systems*, *79*, 12–23.
Xu, X. (2020). Examining an asymmetric effect between online customer reviews emphasis and overall satisfaction determinants. *Journal of Business Research*, *106*, 196–210.
Yang, M., Zheng, Z. (Eric), Mookerjee, V., & Chen, H. (2023). Responding to Online Reviews in Competitive Markets: A Controlled Diffusion Approach. *MIS Quarterly*, *47*(1), 161–194.
Yelp. (2023, April 13). *Yelp announces the 50 most loved brands in the U.S.* Yelp - Official Blog. https://blog.yelp.com/news/yelp-announces-the-50-most-loved-brands-in-the-us/
Yin, D., Bond, S. D., & Zhang, H. (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Quarterly*, *38*(2), 539–560.
Zhang, D., Zhou, L., Kehoe, J. L., & Kilic, I. Y. (2016). What Online Reviewer Behaviors Really Matter? Effects of Verbal and Nonverbal Behaviors on Detection of Fake Online Reviews. *Journal of Management Information Systems*, *33*(2), 456–481.