Marketing Letters
https://doi.org/10.1007/s11002-024-09729-3

The Turing test of online reviews: Can we tell the difference between human-written and GPT-4-written online reviews?

Balázs Kovács[1]

Accepted: 8 April 2024

Abstract
Online reviews serve as a guide for consumer choice. With advancements in large language models (LLMs) and generative AI, the fast and inexpensive creation of human-like text may threaten the feedback function of online reviews if neither readers nor platforms can differentiate between human-written and AI-generated content. In two experiments, we found that humans cannot recognize AI-written reviews. Even with monetary incentives for accuracy, both Type I and Type II errors were common: human reviews were often mistaken for AI-generated reviews, and even more frequently, AI-generated reviews were mistaken for human reviews. This held true across various ratings, emotional tones, review lengths, and participants' genders, education levels, and AI expertise. Younger participants were somewhat better at distinguishing between human and AI reviews. An additional study revealed that current AI detectors were also fooled by AI-generated reviews. We discuss the implications of our findings on trust erosion, manipulation, regulation, consumer behavior, AI detection, market structure, innovation, and review platforms.

Keywords Online reviews · Generative AI · GPT-4 · Turing test

1 Introduction

In our modern world, people can often choose among hundreds and thousands of books to read, restaurants to visit, hotels to stay at, or cat food to buy. People often consult online review websites and read reviews by previous consumers when choosing among these options. It is now well-established that online reviews influence buying decisions and thus drive sales and profits (e.g., Archak et al., 2011; Chevalier & Mayzlin, 2006; Li & Hitt, 2008; Netzer et al., 2012; Pavlou & Dimoka, 2006) (for literature reviews, see Sharkey et al. (2023); Tadelis (2016)). Online reviews typically consist of two

✉ Balázs Kovács
Balazs.kovacs@yale.edu

[1] Yale University, 165 Whitney Avenue, New Haven, CT 06520, USA

Published online: 12 April 2024

components: star ratings and written feedback. Both components can convey a message, and scholars have identified how ratings (Chevalier & Mayzlin, 2006; Dellarocas et al., 2007) and the review text (Archak et al., 2011; Netzer et al., 2012; Pavlou & Dimoka, 2006) can help purchasing decision of future users.

This informative function of online reviews may be seriously threatened with the ascent of generative AI. Recent advancements like GPT-4 can cheaply and quickly generate high-quality texts, potentially threatening the legitimacy of review websites. If readers and AI detectors cannot distinguish whether a given review was written by an AI or by a human, readers may lose their trust in the review platform (Ananthakrishnan et al., 2020; Brandl & Ellis, 2023). Of course, platforms have been putting significant efforts into fighting fake[1] reviews (Han et al., 2022; He et al., 2022; Luca & Zervas, 2016; Pavlou & Gefen, 2004), but many of these methods have relied on the ability of humans and algorithms to recognize fake reviews from the review text. This challenge intensifies if neither humans nor AI detectors can distinguish between human-written and AI-generated reviews.

In this paper, we investigate whether the text created by the currently (March 2024) cutting-edge LLM, GPT-4, can generate restaurant reviews that pass the Turing test (Turing, 1950), which is defined as readers being unable to distinguish whether the reviews were written by a human or an AI. In two experimental studies, we randomly sample restaurant reviews from Yelp.com and we have GPT-4 create similar counterparts for each review (Study 1) or a completely fictional review for the same restaurant (Study 2). We present these reviews to participants and ask them to classify them as human-written or AI-generated. We find that on average people are not able to differentiate between human-written and AI-generated reviews. In an additional study, we also tested if AI-text detectors can tell the difference. Using two AI-text recognition tools, we found that neither could identify GPT-4 generated reviews. Specifically, they misidentify the output of our GPT-4 prompts as human-written text.

These findings have implications for multiple domains and audiences, including the erosion of trust in reviews, potential manipulations by bad actors, regulatory and ethical challenges, impact on marketing and consumer behavior, advancement of AI detection tools, economic impact on businesses, and evolution of consumer review platforms. We discuss these in the second half of the paper, together with recommendations for future research.

2 Brief literature review

2.1 Fake reviews

Fake online reviews are increasingly problematic for review platforms. The "when, why, and who" of fake reviews have gained more scholarly focus (Wu

[1] Zhang et al. (2016) define fake reviews as "deceptive reviews provided with an intention to mislead consumers in their purchase decision making, often by reviewers with little or no actual experience with the products or services being reviewed. Fake reviews can be either unwarranted positive reviews aiming to promote a product, or unjustified false negative comments on competing products in order to damage their reputations."

et al., 2020). Mayzlin et al. (2014) studied hotel reviews on Expedia and Tripadvisor. Their research design utilized the fact that Tripadvisor allows review submission by anyone, whereas Expedia requires actual stay confirmation. They suggest that the difference between the distribution of ratings across platforms indicates fake reviews. They discovered that single-unit-owned independent hotels most likely benefit from posting misleading reviews. They demonstrate that such hotels have more proportionally five-star reviews on Tripadvisor than on Expedia. Luca and Zervas (2016) examined motives for fraudulent Yelp reviews. They report that about 16% of Yelp's restaurant reviews are flagged as fraudulent. They observed that these fraudulent reviews often exhibit more extreme sentiments compared to genuine reviews. The study further reveals that independent and less reputable restaurants are likelier to commit review fraud. He et al. (2022) studied counterfeit product reviews on Amazon.com. Utilizing data from an illicit market selling fake reviews, they found that acquiring such reviews leads to a temporary yet noticeable boost in both average rating and review volume. However, once companies cease purchasing fake reviews, average ratings fall, and one-star reviews significantly increase. Wu et al. (2020) show that fake reviews mainly aim to boost brand image and harm competitors, and are common with lower-quality firms and early in a product's life cycle.

Fake reviews negatively impact review platforms[2] and society by reducing online review usefulness and quality (Wu et al., 2020; T. Zhang et al., 2017) and by decreasing product-to-customer matching efficiency (Mayzlin et al., 2014). Fake reviews mislead consumers (Zhao et al., 2013), damage review credibility (Luca & Zervas, 2016), decrease review helpfulness (Agnihotri & Bhattacharya, 2016), and weaken purchase intentions (Ahmad & Sun, 2018).

Review websites employ several strategies to filter out fake reviews, aiming to maintain their platforms' credibility and usefulness. There are three main types of strategies. One that does not rely on analyzing review texts, and two that do. The first group contains detection algorithms that analyze patterns in review submissions and flag if unusually many reviews come from a given IP address or flag a sudden surge of positive or negative reviews (Dellarocas, 2003). Platforms also may cross-reference review data with other information, such as user account activity or purchase history, and may require users to verify their identity through email or text message (Laudon & Laudon, 2004). The second group encompasses automatic text analysis tools that identify red flags, including unnatural language, repetitive phrasing, specificity of the review, use of certain words, or overly biased reviews (Han et al., 2022; Mudambi & Schuff, 2010; Wu et al., 2020). The third group of tools assumes that humans can tell AI-generated text. These include approaches where human moderators and users can flag suspicious reviews for further inspection (Cheung & Lee, 2012; Kozinets, 2002).

[2] Wu et al. (2020) highlight an interesting exception: some newly established review platforms intentionally add fake reviews and copy reviews from other platforms to give the impression that their platform is widely used, thereby circumventing the catch-22 of platforms: users do not arrive until reviews are posted, and reviews are not posted until users arrive.

As we will demonstrate in this paper, these latter two categories of tools are becoming obsolete with the advent of GPT-4.

2.2 Detecting human- vs AI-generated text

While there is currently no study exploring the distinction between human-written and GPT-4-generated reviews, related research with earlier AI models and non-review texts has documented the limitations of AI-text recognition, both by humans and algorithms. A 2019 study showed that shorter AI text can fool readers (Ippolito et al., 2019). Köbis and Mossink (2021) found that humans could distinguish human-written poems from GPT-2-generated ones only if the former were written by professional poets; otherwise, the difference was indiscernible. Recent studies in professional, hospitality, and dating contexts (Jakesch et al., 2023) and in health, finance, entertainment, technology, and travel contexts (Brandl & Ellis, 2023) demonstrated that humans were not able to tell human-written vs AI-generated content. Additionally, this research indicated that individuals with greater familiarity with AI were marginally more successful at identifying AI-generated content.

We contribute to this literature by exploring whether people can tell human-written vs AI-generated reviews. Focusing on online reviews is revealing because they encompass various styles and are typically written in a non-professional manner, which could complicate AI generation. Moreover, the brevity of reviews further complicates detection (Ippolito et al., 2019). Earlier studies used older methods; GPT-4's advancements may render these findings outdated (Le Mens et al., 2023).

3 Study 1

3.1 Online reviews from Yelp

To create a sample of real human-generated reviews, we sampled restaurant reviews from the publicly available Yelp review dataset (2022 version), which contains approximately 7 million reviews about the businesses in North American metropolitan areas such as Philadelphia, Pittsburgh, or New Orleans. We decided to use this data source as it is commonly used in the marketing (e.g., Pentina et al., 2018), sociology (Kovács, 2024), and management literatures (e.g., Kovács & Carroll, 2023; Kovács et al., 2014) to study online reviewing behavior. We analyzed reviews from 2019 to ensure comparability with pre-COVID and pre-ChatGPT periods, making them unlikely to be AI-generated due to the less sophisticated LLMs available at that time. Moreover, we selected from reviews that were not tagged as fake or spam by Yelp's algorithm.

We took a random stratified sample of restaurant reviews from 2019, such that we took 20 reviews for each possible star rating level, e.g., 1 star, 2 stars, …, 5 stars.

Table 1 Examples of human-contributed Yelp reviews and their AI-generated counterparts (these are actual stimuli used in the experiment)

Human-contributed Yelp review	AI-generated counterpart
Worst FSU game watching site. DJ would rather play crap music vs have the sound on for the game. If you're going to commit to FSU for games, especially as an Alum, get your staff right	Terrible spot for FSU games! The DJ's obsessed with his tunes, ignoring the game. Why bother being an FSU bar if you won't play the game sound? Alumni like me, we expect better—train your people!
Great diner for breakfast & lunch! My kids love the krispy pancakes and hash Browns. We use to get meatball sandwich all the time, our favorite but it's no longer on the menu. Please bring it back!	Love this spot for brunch vibes! The kiddos devour those crispy pancakes & the hashies. Totally miss their meatball sub though – was our go-to! Wish they'd bring it back ASAP!!
The Blind Onion is one of our favorites. Excellent pizza crust and yummy wings!! Cheesy bread is very yummy too!	Absolutely adore the Blind Onion! Their pizza dough? TO DIE FOR – crispy and just perfect. Wings are just delish, every single time. And don't even get me started on the cheezy bread – A MUST! 10/10 would recommend!!

The result is a sample of 100 restaurant reviews. Table 1 shows three (randomly chosen) Yelp reviews used as a stimulus.

3.2 GPT-4 generated reviews

To generate AI-generated reviews, we fed into OpenAI's API (GPT-4 Turbo model, accessed November 29, 2023) the following prompt, where *[review_text]* stands for the actual review text.

> "Here is an online review of a restaurant. '[review_text]' Write a fictional Yelp online review about the same restaurant (keep the name), about the same length, same tone, same specificity, same emotionality, same style (ok to have some typos), touching on similar topics. Make it MUCH shorter. Write as if you were a human. Make it as short as the original text. Add two typos! Feel free to use colloquial spelling and sometimes even ALL CAPS when emphasizing something."[3]

We saved the outcomes of these queries, creating an AI-generated counterpart for each actual Yelp review. See Table 1 for examples.

Importantly, we wish to emphasize that our objective is not to prove that *any* computer-generated review could compare to human-written reviews. Instead, our aim is to provide an existence proof and explore whether a widely accessible generative AI, using a simple prompt, can produce human-like reviews.

[3] This refined prompt is based on a simpler one from our pilot study, where we found that GPT-4 produces longer texts without shortening instructions. Participants often identified human-generated reviews by typos, misspellings, or unusual spellings like ALL CAPS, leading us to incorporate these in the GPT prompt.

Table 2 Cross-classification table of human- vs AI-generated reviews and their classification by experiment participants (Study 1)

	Participants classified it as human-written	Participants classified it as AI-written	Total
Human-written Yelp review	907	604	1,511
AI-generated review	896	611	1,507
Total	1,803	1,215	3,018

3.3 Method and participants

Participants were informed that we were studying whether humans can distinguish between human-written and AI-generated restaurant reviews. Each participant was shown 20 random reviews from a sample of 200 (100 real, 100 AI-generated).[4] The experiment involved no deception. For each review, they had to choose between AI- or human-generated (binary forced choice, no scale). After classifying the reviews, participants were asked to provide demographic information. On a separate screen, they could answer in an essay box about the criteria they used to distinguish between human- and AI-generated reviews.

We recruited 151[5] participants through Prolific Academic (55% female, average age of 47 years). The selection criteria required participants to be adults, native English speakers, and residents of a majority English-speaking country (USA, Canada, UK, Ireland, Australia). To ensure participants paid attention and took the task seriously, they were promised an extra $0.50 bonus (in addition to the $1.50 base payment) for correctly identifying at least 16 out of the 20 reviews as human- or AI-generated. Only 6 out of the 151 participants managed to do so.

Overall, our data consist of 3018 classifications of review text to human- vs AI-generated. The experiment was approved by the IRB Board of Yale University and is pre-registered at https://osf.io/38qtz.

3.4 Study 1 results

Table 2 presents the main results of Study 1: the cross-tabulation of human- vs AI-generated reviews and their classification by experiment participants. The table clearly demonstrates that participants cannot distinguish between reviews written by humans and those generated by AI. If participants were able to correctly classify reviews as human-written or AI-generated, we would see zeros in the off-diagonals (which show the misclassified reviews). Instead, this pattern resembles a coin toss: approximately half of the reviews classified by participants as written by humans are

[4] Given the full randomization, participants may or may not have seen both a human- and an AI-written review of the same restaurant.

[5] We targeted 150 participants, but after a participant timeout and replacement by Prolific, the original participant returned, completing the survey, and resulting in 151 participants.

actually AI-generated, and half of those classified as AI-generated are, in fact, written by humans. The chi-square statistic with Yates correction is 0.0798 ($p = 0.777$), indicating that there is no statistically significant relationship between whether a review was human-written and whether participants recognize it as such. Therefore, GPT-4, with our specific prompt, passes the Turing test for review generation.

We examined potential moderating effects of reader characteristics (gender, age, education), review characteristics (length, valence, emotionality, presence of typos, profanity, and informal expressions),[6] and readers' efforts in answering the questions (proxied by the amount of time spent reading the reviews). Table 3 shows the results. In these logit models, the dependent variable is whether the participant correctly identifies the review as human-written or AI-contributed. Model 1 is estimated on the full sample, while models 2 and 3 are estimated on the subsample of human-written (model 2) and AI-generated (model 3) reviews to allow for separating Type I and Type II errors. The only significant effect worth noting is that younger participants are only slightly more adept at differentiating between human and AI-written reviews.

4 Study 2

Study 2 was designed to overcome some of the limitations of Study 1 and to explore the robustness of the results. The study was based on Study 1's design, but we made four major modifications. First, instead of creating an AI-modified version of an existing review, we randomly selected 100 restaurants,[7] and, for each, randomly selected 10 reviews. We then asked GPT-4 to create an entirely fictional review of the same restaurant based on the 10 selected reviews, without reusing any language from the existing reviews. Second, to test whether the age effect identified in Study 1 resulted from differences in participants' AI-literacy, we collected data on their experiences with AI chatbots. Third, because the binary AI-written/human-written response design in Study 1 did not permit a detailed understanding of participants' decision-making between AI-written and human-written reviews, we revised the answer format in Study 2 to a 5-item Likert scale. Specifically, for each displayed review text, we asked, "Is this restaurant review written by a human or by an AI?" with the following options: "Most likely human," "Probably human," "Can't decide," "Probably AI," "Most likely AI." Fourth, we introduced an attention check; in one of the review texts, we included the message, "This is an attention check. If

[6] We used ChatGPT to code the reviews for valence, emotionality, presence of typos, profanity, and informal expressions. Specifically, we instructed GPT-4 to "Here is a restaurant review. [XXX] Code this review for each of the following dimensions: sentiment (from 0 to 100, where 100 is highly positive), emotionality (from 0 to 100, where 100 is highly sentimental), the number of typos or misspellings, the number of profane words or expressions, and the number of informal expressions. Put in a table." We cross-checked a few of these answers and agreed with GPT-4's answers so we used these values in these regressions.

[7] The sample of restaurants in Study 2 is different from the sample of restaurants and reviews in Study 1. In Study 2, we only included restaurants that received at least 10 English-language reviews in 2019.

Table 3 Determinants of correct classification of human-written vs AI-generated reviews (Study 1)

DV: correct classification	(1)	(2)	(3)
Sample	All reviews	Human-written reviews	AI-written reviews
Female participant	0.0811	0.195	−0.0256
	(1.09)	(1.81)	(−0.24)
Education[+]	−0.0258	−0.0291	0.000333
	(−1.00)	(−0.79)	(0.01)
Age[++]	−0.112***	−0.115**	−0.107*
	(−3.84)	(−2.75)	(−2.49)
Review length (in characters)	0.000351	−0.000741	−0.00105
	(1.04)	(−1.41)	(−1.88)
Duration (in seconds)	0.000179	0.000220	0.000199
	(1.34)	(1.16)	(1.01)
Emotional intensity of rating (0–100)	0.00478*	0.00430	−0.0000180
	(2.06)	(1.28)	(−0.01)
Positivity of rating (0–100)	0.000776	−0.00904***	0.0103***
	(0.68)	(−5.20)	(6.35)
Vulgar words #	0.250	−0.00551	−0.743*
	(1.56)	(−0.02)	(−2.20)
Informal words #	−0.133***	0.0824	0.0740*
	(−7.49)	(1.33)	(2.15)
Misspelled words #	−0.126	−0.221	0.298
	(−0.69)	(−0.80)	(1.25)
Constant	0.397	1.188***	−0.523
	(1.70)	(3.61)	(−1.47)
N	3018	1511	1507

[+]Education is coded as 1 (some high school or less), 2 (high school or GED), 3 (some college but no degree), 4 (associates or technical degree), 5 (Bachelor's), 6 (Graduate or professional degree)

[++]Age is coded as 2 (18–24 yo), 3 (25–34 yo), 4 (35–44), 5 (45–54), 6 (55–64), and 7 (65 +)

Robust standard errors. *T*-statistics in parentheses

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

you are reading this, select "Most likely human." Other than these modifications, the design of Study 2 was identical to that of Study 1.

4.1 Method and participants

As in Study 1, we recruited 150 participants through Prolific Academic (58% female, average age 47.5 years), all native English speakers and residents of majority English-speaking countries (US, Canada, UK, Ireland, Australia). To ensure participants' attentiveness and seriousness, they were promised an additional $0.50 bonus (beyond the $1.50 base payment) for correctly identifying at least 16 of the 20 reviews as human- or AI-generated. Only one person missed

Table 4 Cross-classification table of human- vs AI-generated reviews and their classification by experiment participants (Study 2)

	Participants classified it as					Total
	Most likely human	Probably human	Can't decide	Probably AI	Most likely AI	
Human-written Yelp review	369	434	76	350	281	1510
AI-generated review	391	506	85	344	160	1486
Total	760	940	161	694	441	2996

the attention check, so we present the results with the full sample (results are almost identical when we exclude the person who missed the attention check).

4.2 Results of Study 2

As in Study 1, we first present the results in a cross-tabulation format (Table 4). The table clearly illustrates that participants struggle to distinguish between human-written and AI-written reviews. Out of the 1510 times they were presented with an actual human-written restaurant review, only 803 times (53.2%) did they think that the review was most likely or probably human-written. In 41.8% of cases, participants mistook the human-written reviews for AI-written ones. Interestingly, they performed even worse when judging AI-written reviews: out of the 1486 times they were presented with an AI-written review, 64% of the time they thought that the review was human-written. Only in 34% of the cases did participants correctly identify the review as AI-written. Moreover, participants were not only unsure whether the reviews were written by humans but in many cases, they were also confident they were human-written, even when this was not the case. Interestingly, there is a slight asymmetry in the responses, with participants appearing more confident in their judgment of reviews they perceive as human-written, even when incorrect, compared to those they consider AI-written. These patterns, as the results of Study 1, indicate that the AI-written reviews pass the Turing test. Even more than in Study 2, the results here, showing that AI-written reviews are perceived as more human than human-written reviews, indicate a situation that AI researchers term "AI hyperrealism" (Miller et al., 2023).

In Study 2, as in Study 1, we investigate the factors that may influence whether participants can correctly tell human-written reviews from AI-written reviews. Because in this study, participants answered the question about human-vs-AI authorship in a 5-item Likert scale; we first dichotomized the answers. Specifically, a classification is deemed correct if the participant classified a human-written review as "most likely human" or "probably human" or an AI-written review as "most likely AI" or "probably AI." As in Study 1, we run logit regressions. Table 5 shows the results. Model 1 is estimated on the full sample, while models 2 and 3 are estimated on the subsample of human-written (model 2) and

Table 5 Determinants of correct classification of human-written vs AI-generated reviews (Study 2)

DV: correct classification	(1)	(2)	(3)
Sample	All reviews	Human-written reviews	AI-written reviews
Female participant	−0.0983	−0.0950	−0.109
	(−1.24)	(−0.85)	(−0.93)
Education[+]	0.00240	−0.0370	0.0516
	(0.09)	(−1.03)	(1.36)
Age[++]	−0.0895**	−0.0969*	−0.0903*
	(−2.99)	(−2.32)	(−2.03)
Experience with generative AI	−0.0720	−0.0423	−0.120
	(−1.39)	(−0.59)	(−1.55)
Review length (in characters)	−0.000661	−0.00149**	0.00186
	(−1.26)	(−2.58)	(1.34)
Duration	−0.0000782	0.000246	−0.000443
	(−0.38)	(0.82)	(−1.46)
Emotional intensity of rating (0–100)	−0.000973	0.00286	0.0105
	(−0.29)	(0.70)	(1.00)
Positivity of rating (0–100)	−0.00388*	−0.00745***	0.0103***
	(−2.43)	(−3.33)	(3.41)
Informal words #	−0.208***	−0.0645	−0.0211
	(−6.08)	(−1.13)	(−0.33)
Misspelled words #	0.719***	0.545***	0.000
	(4.53)	(3.41)	(0.000)
Constant	1.155***	1.325**	−1.907*
	(3.52)	(3.14)	(−2.06)
N	2976	1501	1475

[+]Education is coded as 1 (some high school or less), 2 (high school or GED), 3 (some college but no degree), 4 (associates or technical degree), 5 (Bachelor's), 6 (Graduate or professional degree)

[++]Age is coded as 2 (18–24 yo), 3 (25–34 yo), 4 (35–44), 5 (45–54), 6 (55–64), and 7 (65 +)

Robust standard errors. *T*-statistics in parentheses

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

AI-generated (model 3) reviews to allow for separating Type I and Type II errors. The model specifications build upon those of Study 1 and include the newly collected "AI-expertise" variable assigned a value of 1 if the participant has experience using generative AI models (0 otherwise). As in Study 1, we find that most of the factors have insignificant or inconsistent effect on whether participants can correctly identify the authorship of a review. The only consistent finding still is the negative age effect, i.e., that younger participants are better in identifying AI-written reviews. Interestingly, this effect cannot be explained by differences in participants' experience with generative AI models, as that variable is insignificant in the models.

4.3 Limitations of Studies 1 and 2 and potential for future research

Studies 1 and 2 offer insights into distinguishing human-written from GPT-4 generated reviews, with limitations suggesting avenues for future research. Expanding the scope beyond restaurant reviews could reveal more about the application and perception of AI-generated content. For example, using non-review texts as stimuli, Brandl and Ellis (2023) demonstrated varying abilities in detecting AI-generated text across domains, with the best detection in tech and the worst in travel (the difference was only 10% though). In addition, investigating reviews in different languages may also uncover cultural or linguistic nuances in the detection of AI-generated text.

Another area to explore is varying the real vs fake review ratio and informing participants about it. Exploring this could reveal how awareness of AI influences judgment. Additional information, such as previous reviews or profile pictures, might deepen the understanding of AI-content detection. Furthermore, adding a third category of "human plus AI"-generated content could also yield interesting results.

Future research could also explore variation across review content. For example, Jago (2019) found that descriptions involving moral authenticity are more penalized when AI-written compared to those by humans, while humans are more tolerant of AI-generated content regarding type authenticity. Similarly, there may be better differentiation between human and AI reviews on moral vs procedural issues (e.g., the price level or opening hours of a restaurant).

Furthermore, while participants in Study 1 could not differentiate between human- and AI-generated reviews, Study 2 indicated a hyperrealism wherein AI-generated reviews were perceived as more human-like than those generated by humans. Future research should further explore this difference and test whether it represents a systematic effect attributable to some of the changes implemented from Study 1 to Study 2, such as the different methods of generating reviews or the altered response format. (The attention check and the AI-expertise question cannot cause this change as they were presented after participants answered the Turing test questions.)

Finally, it could be argued that demonstrating a null result, like ours, is challenging, as achieving statistical significance might be feasible with a larger sample. Even if that is the case, we maintain that the size of the effect is minuscule; therefore, even if the effect were to become statistically significant, it would not constitute a practically meaningful or useful effect.

5 Study 3: Can AI detectors tell if a review text is AI-written?

Before discussing the implications of our main findings, we examine whether current AI detection tools can differentiate between human- and AI-contributed reviews. The answer to this question has important implications for our main findings.

The algorithmic recognition of human- vs AI-generated text is currently a hot and contentious topic. In the last few years, multiple approaches have been proposed and tried, such as the use of large language detection models, statistical and

stylometry analysis of text, contextual analysis, or watermarking (for a review, see Uchendu et al. (2021)). GPT-4 creates text that often deceives detectors, particularly through intentional typos and colloquialisms, as specified in our prompt. AI detectors struggle with short texts, which are common in online reviews. Revealingly, OpenAI, the creator of GPT-4 released on January 31, 2023, an AI classifier for indicating AI-written text, but later discontinued the service saying "the AI classifier is no longer available due to its low rate of accuracy." https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

To explore whether current AI detectors work for online reviews specifically, we tested two approaches. First, we fed back the 200 reviews from Study 1 to GPT-4, and for each review, asked GPT-4 the following prompt:

> What is the likelihood that this online review was written by an AI program vs a human? Answer on a scale of 0 to 100 where 100 stands for 'for sure AI'. Just give me the number.

We conducted a *t*-test to determine if AI-generated reviews scored higher than human-written ones. No significant effect was found ($t = 0.944$; $p = 0.17$), suggesting that it is currently impossible for GPT-4 to discern whether these reviews are human- or AI-written. Most answers fell in the 10–20 range, indicating that AI-generated reviews mimic human writing and that some human reviews appear AI-like.

Second, we fed the reviews in our sample to the currently most accurate publicly available AI-text recognition software, Copyleaks (Orenstrakh et al., 2023). Because Copyleaks only takes texts that are at least 300 characters long, we could only analyze 102 of our 200 reviews. Copyleaks could not distinguish AI-generated text either, labeling all submitted reviews as human-generated.

6 Implications

The finding that large language models (LLMs) can cheaply and quickly generate online review texts indistinguishable from those written by humans has wide-ranging implications.

6.1 Erosion of trust in reviews

A BrightLocal survey shows that 87% of consumers read online reviews for local businesses and 79% consider them as trustworthy as personal recommendations (www.brightlocal.com/research/local-consumer-review-survey-2020/). Once consumers understand the ability of LLMs to generate authentic-looking reviews quickly and cheaply, it will likely lead them to second-guess whether a review was written by a person or an AI. Such skepticism could damage the reputation and user engagement of review platforms and may affect consumers' purchasing decisions.

6.2 Manipulation and bias

Businesses might use LLMs to create positive reviews for themselves or negative reviews for competitors (Luca & Zervas, 2016; Mayzlin et al., 2014). This could lead to an unfair market, where the quality of products and services is overshadowed by the ability to manipulate online perceptions. Even more so, businesses may contract external services that are set up with numerous accounts to post a large number of fraudulent reviews (He et al., 2022). While these options have been available to firms since the dawn of online reviewing, the situation now changes substantially because humans will not be required to write authentic-looking reviews, making the production of fake reviews much quicker and cheaper.

6.3 Impact on marketing and consumer behavior

AI's ability to create convincing reviews means marketers must adjust their strategies. Traditional marketing focused on positive branding and authentic testimonials. With AI reviews, there may be a shift to authenticating customer experiences. Marketers could focus more on campaigns displaying real interactions and verifiable feedback, such as video testimonials.

For consumers, reduced trust in online reviews might affect purchasing decisions. As trust in written reviews erodes, consumers may turn to word-of-mouth recommendations or direct experiences with products and services. This could lead to a renewed emphasis on personal networks and community recommendations.

Flooding review sites with fake positive reviews can create a false sense of consensus, where people are influenced by the opinions of others in their behavior. This resembles "astroturfing," where organizations create an illusion of grassroots support for a product or service. However, AI elevates this manipulation to a new level, due to its scale and believability.

6.4 Asymmetric economic costs for small business

The economic impact of AI-generated online reviews can be significant, especially for small businesses. Larger companies have more ad budgets, but small businesses, like family restaurants, depend on online reviews for visibility (Luca & Zervas, 2016; Mayzlin et al., 2014). If AI reviews become indistinguishable, it could harm small businesses that rely on genuine reviews. Consequently, they may need to increase marketing and outreach efforts to combat false AI reviews, potentially hindering new businesses and limiting competition.

6.5 Regulatory and ethical challenges

The advent of AI-generated reviews indistinguishable from those written by humans poses significant regulatory challenges. Regulators need to ensure transparency and fairness in online marketplaces, possibly through AI authorship

disclosure regulations. The Federal Trade Commission (FTC) in the US has guidelines for endorsements in advertising, which could extend to AI content. These guidelines require disclosing connections like payments or free products between endorsers and advertisers. A similar principle could mandate AI-generated review disclosure and labeling to maintain transparency.

6.6 Evolution of consumer review platforms

Review platforms currently use various anti-spam strategies. These strategies ensure platform credibility and usefulness. Some of these processes are not directly impacted by the ability of new LLMs to create text indistinguishable from human-generated text. This includes algorithms that flag numerous reviews from the same IP address or unusual review surges, indicating potential manipulation (Dellarocas, 2003). Platforms might verify reviews by cross-referencing with account activity or purchase history (Laudon & Laudon, 2004). Additionally, platforms frequently require email or phone verification (Pavlou & Gefen, 2004).

Other ways to detect fake reviews are becoming ineffective. These include text analysis tools targeting red flags like unnatural language or extreme sentiment (Han et al., 2022) and filters that consider feedback specificity, word use, or tone (Mudambi & Schuff, 2010). In addition to automated tools, some other tools were built on the assumption that humans can tell AI-generated text. These involve human moderators and community reporting for suspicious reviews (Cheung & Lee, 2012; Kozinets, 2002). As our results indicate, such tools are quickly becoming ineffective.

Indistinguishable human-AI reviews require rethinking platform authenticity mechanisms. Platforms could enforce stringent review verification, like proof of purchase or experience (Amazon's "Verified Purchase"), or alternative proofs (restaurant receipts on Yelp, hotel reservations on Tripadvisor). Additionally, platforms should inform users about AI content, possibly by labeling AI-generated reviews (Ananthakrishnan et al., 2020).

7 Conclusion

In summary, our findings that neither humans nor AI can tell human-generated reviews from AI-generated reviews indicate that the online review domain is likely to undergo substantial changes. Consumer and producer behaviors will evolve, causing trust issues and market distortions. Consumer review platforms, along with potential regulatory forces, need to respond to this with a combination of technological solutions, policy changes, and user education efforts.

Acknowledgements This research has benefitted from feedback from Glenn Carroll, Jennifer Dannals, Jerker Denrell, Balázs Gyenis, Arthur Jago, and Iris Wang. All remaining errors are my own.

Data availability Data and code is available from the author at request.

Declarations

Ethical approval Yale University's IRB Board approved the research, IRB # 1508016387.

Informed consent Consent was collected at the beginning of the experiment.

Conflict of interest The author declares no competing interests.

References

Agnihotri, A., & Bhattacharya, S. (2016). Online review helpfulness: Role of qualitative factors. *Psychology & Marketing, 33*(11), 1006–1017.

Ahmad, W., & Sun, J. (2018). Modeling consumer distrust of online hotel reviews. *International Journal of Hospitality Management, 71*, 77–90.

Ananthakrishnan, U. M., Li, B., & Smith, M. D. (2020). A tangled web: Should online review portals display fraudulent reviews? *Information Systems Research, 31*(3), 950–971.

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.

Brandl, R., & Ellis, C. (2023). Survey: ChatGPT and AI Content –Can people tell the difference? Retrieved from https://www.tooltester.com/en/blog/chatgpt-survey-can-people-tell-the-difference/

Cheung, C. M., & Lee, M. K. (2012). What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. *Decision Support Systems, 53*(1), 218–225.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*(3), 345–354.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49*(10), 1407–1424.

Dellarocas, C., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing, 21*(4), 23–45.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: Concepts and techniques*. Morgan Kaufmann.

He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. *Marketing Science, 41*(5), 896–921.

Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2019). Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint* arXiv:1911.00650

Jago, A. S. (2019). Algorithms and authenticity. *Academy of Management Discoveries, 5*(1), 38–56.

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences, 120*(11), e2208839120.

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior, 114*, 106553.

Kovács, B. (2024). Studying travel networks using establishment Covisit networks in online review data. *Socius, 10*, 23780231241228916.

Kovács, B., & Carroll, G. R. (2023). Distinguishing between cosmopolitans and omnivores in organizational audiences. *Academy of Management Discoveries, 9*(4), 549–577.

Kovács, B., Carroll, G. R., & Lehman, D. W. (2014). Authenticity and consumer value ratings: Empirical tests from the restaurant domain. *Organization Science, 25*(2), 458–478.

Kozinets, R. V. (2002). The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research, 39*(1), 61–72.

Laudon, K. C., & Laudon, J. P. (2004). *Management information systems: Managing the digital firm*. Pearson Education.

Le Mens, G., Kovács, B., Hannan, M. T., & Pros, G. (2023). Uncovering the semantics of concepts using GPT-4. *Proceedings of the National Academy of Sciences, 120*(49), e2309350120.

Li, X., & Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research, 19*(4), 456–474.
Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science, 62*(12), 3412–3427.
Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review, 104*(8), 2421–2455.
Miller, E. J., Steward, B. A., Witkower, Z., Sutherland, C. A., Krumhuber, E. G., & Dawel, A. (2023). AI hyperrealism: Why AI faces are perceived as more real than human ones. *Psychological Science, 34*(12), 1390–1403.
Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon.com. *MIS Quarterly, 34*(1), 185–200.
Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science, 31*(3), 521–543.
Orenstrakh, M. S., Karnalim, O., Suarez, C. A., & Liut, M. (2023). Detecting llm-generated text in computing education: A comparative study for chatgpt cases. *arXiv preprint* arXiv:2307.07411
Pavlou, P. A., & Dimoka, A. (2006). The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Information Systems Research, 17*(4), 392–414.
Pavlou, P. A., & Gefen, D. (2004). Building effective online marketplaces with institution-based trust. *Information Systems Research, 15*(1), 37–59.
Pentina, I., Bailey, A. A., & Zhang, L. (2018). Exploring effects of source similarity, message valence, and receiver regulatory focus on yelp review persuasiveness and purchase intentions. *Journal of Marketing Communications, 24*(2), 125–145.
Sharkey, A., Kovács, B., & Hsu, G. (2023). Expert critics, rankings, and review aggregators: The changing nature of intermediation and the rise of markets with multiple intermediaries. *Academy of Management Annals, 17*(1), 1–36.
Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics, 8*, 321–340.
Turing, A. M. (1950). Computing machinery and intelligence. *Mind, LIX*(236), 433–460.
Uchendu, A., Ma, Z., Le, T., Zhang, R., & Lee, D. (2021). Turingbench: A benchmark environment for Turing test in the age of neural text generation. *arXiv preprint* arXiv:2109.13296
Wu, Y., Ngai, E. W., Wu, P., & Wu, C. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems, 132*, 113280.
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.
Zhang, T., Li, G., Cheng, T., & Lai, K. K. (2017). Welfare economics of review information: Implications for the online selling platform owner. *International Journal of Production Economics, 184*, 69–79.
Zhao, Y., Yang, S., Narayan, V., & Zhao, Y. (2013). Modeling consumer learning from online product reviews. *Marketing Science, 32*(1), 153–169.