Generative AI and the Perceived Quality of User-Generated Content: Evidence from Online Reviews

Samsun Knight, University of Toronto*
Yakov Bart, Northeastern University†
Minwen Yang, University of Toronto‡

Date: November 4, 2025

Abstract

How does generative AI use affect the perceived quality of online reviews? Our large-scale analysis of recent restaurant reviews on Yelp.com and product reviews on Amazon.com shows that detected use of generative AI in review writing is associated with significant declines in perceived quality. These results are similar both in OLS and in two-period differences-in-differences estimation based on within-reviewer changes in AI use. Moreover, we document that the results are driven by reviews without expert reviewer badges. Three pre-registered studies recreate this effect from the field and further establish that 1) simply informing readers that a review is written by AI, regardless of whether it actually is, results in lower perceived review quality, and 2) this effect largely disappears when reviews are described as based on verified customer experiences, suggesting that suspicion of fake reviews is a primary driver of the observed effect.

KEYWORDS: generative artificial intelligence, ChatGPT, online reviews, user-generated content

*samsundknight@gmail.com
†y.bart@northeastern.edu
‡minwen.yang@rotman.utoronto.ca

We thank participants of the Cornell ESIF: Economics and AI+ML conference, the Wharton Business & Generative AI Workshop, the Yale SOM Conference on Artificial Intelligence, Machine Learning and Business Analytics conference, and the Northeastern brown bag series for helpful comments and feedback.

1 Introduction

Online reviews and ratings are a core part of the internet for contemporary consumers, allowing them to freely access crowdsourced information on a wide variety of products and offerings that would otherwise be costly to learn independently. This online review ecosystem is essential to many major internet commerce platforms and companies, and also provides broader social surplus by way of an improved information environment and resulting efficiency gains (Reimers and Waldfogel 2021, Chevalier and Mayzlin 2006, Babić Rosario et al. 2016). Importantly, this ecosystem relies both on the content itself, which determines how useful the information is, as well as consumer perceptions of that content, which determines how *used* that information is in decision-making (Mudambi and Schuff 2010).

At the same time, the production process for user-generated online reviews is currently undergoing a massive paradigm shift following the recent release of ChatGPT and other generative AI tools, as widely available generative AI chatbots offer brand-new ways for reviewers to produce content. On the one hand, some observers argue that easier access to generative AI will yield higher-quality reviews, particularly from users who struggle with grammar or fluency (Noy and Zhang 2023, Li et al. 2023b). On the other hand, there are concerns that generative AI usage may lead to low-effort review production that may deteriorate overall information quality and lead consumers to trust online information less (Castelo et al. 2019, Longoni et al. 2019, Brynjolfsson et al. 2023, Doshi and Hauser 2023, Roose 2023). A meaningful deterioration in trust, in particular, could lead to significant losses for review platforms and affected businesses (as well as efficiency losses for society as a whole) if consumers become less inclined to rely on otherwise-useful information simply because they perceive it to be more unreliable, even if the average information value of content is in fact largely unchanged or improved.

In this paper, we present novel empirical evidence that usage of generative AI is associated with significantly lower perceived review quality, both in two large-scale field studies and in three pre-registered online experiments. Our evidence suggests that, while positive effects may be possible from responsible generative AI usage, negative effects associated with AI-written product reviews currently predominate in the environments that we inspect. This is reflected in lower perceived review quality,

defined as consumers' subjective judgments of characteristics such as helpfulness, usefulness, and persuasiveness.

First, we investigate field data from two prominent websites that feature online reviews, Yelp.com and Amazon.com, using samples of hundreds of thousands of reviews written before and after ChatGPT's public release. On both of these major online review websites, we find that detected use of generative AI is associated with significantly lower perceived quality of content as measured by other users' ratings of the usefulness (or helpfulness) of reviews. [1] For Yelp.com, detected AI usage in the review production process is associated with a 16.6% standard deviation decrease in the number of "useful" votes, while for Amazon.com, detected AI usage is associated with a 14.2% standard deviation decrease in the number of "helpful" votes. This result holds across a variety of robustness checks and alternative specifications, including those that regress against different measures of review quality, specifications with and without a rich set of controls, and even a stringent two-period differences-in-differences estimation that identifies the effect from only within-reviewer changes between the pre-period before ChatGPT's public release (January 2022-November 2022) and the post-period (December 2022-September 2023). This differences-in-differences model controls for both potential selection into ChatGPT adoption and potential associations between review recency and accumulated ratings through author and time period fixed effects, respectively. In this specification, comparing quality changes within-reviewers across the pre- and post-release period, we find that a standard deviation increase in AI adoption leads to a highly significant 6.5% standard deviation decrease in useful votes on Yelp, compared to null effects in the pre-period.[2]

We also find evidence of significant heterogeneity in this deleterious effect of detected AI usage, associated with the presence of expert reviewer badges. Splitting our sample between reviews with expert status badges (written by "Elite" reviewers on Yelp, or "Vine Voice" reviewers on Amazon) and those without, we find that this negative review quality effect appears to be driven entirely by non-badged reviews in

[1] We use ZeroGPT, an online provider of generative AI detection software, to classify reviews as detected to involve AI usage or not; while this ZeroGPT-detected usage may not correspond exactly with actual usage, in a subsequent online experiment we show that detected AI usage significantly predicts participant perceptions of AI usage.

[2] Coefficient estimates for differences-in-differences model on our Amazon.com sample are similar, although we have a far smaller matched sample of tracked reviewers in both periods for that sample, and results are not statistically significant.

our differences-in-differences specification: non-badged reviews account for the entire observed effect for Yelp.com, with a tightly estimated zero effect for badged reviews, and point estimates are approximately twice as large for non-badged versus badged reviews in our Amazon.com sample.

While this field evidence suggests a strong negative effect of AI usage, one may still be concerned that there are further confounds or spurious artifacts of the data that we are not able to observe or control for in our empirical specifications. Moreover, our field results may be driven by some combination of effects of actual AI usage and effects of perceived AI usage, but in our observational panels we are unable to separate these two potential effects, as perceptions of AI usage and actual AI usage are likely to be highly collinear.

To alleviate such concerns, we conducted three controlled online experiments. In our first pre-registered study, we decompose the effect between effects of perceptions of AI usage and effects of actual AI usage, either of which may plausibly explain the patterns we find in our field data. We present participants with reviews that are either 1) explicitly described as AI-written and are actually AI-written, 2) described as human-written and actually AI-written, 3) described as AI-written and actually human-written, or 4) described as human-written and actually human-written, and then ask participants to rate each such review across a set of perceived quality dimensions. Comparing these conditions, we find a strong negative effect on perceived quality of described AI use, but no effect of actual AI use: identical sets of product reviews, when described as generated by ChatGPT, are rated as significantly lower-quality across a wide set of dimensions, including usefulness, helpfulness, and persuasiveness. However, when told that reviews are human-written, participants do not perceive AI-written reviews as lower-quality compared to human-written ones. This evidence suggests that the negative penalty associated with AI usage stems from whether a product review is perceived to be written by AI, rather than whether it was actually written by AI or not.

We dig further into this perception-channel result with a second pre-registered study, which seeks to determine the extent to which this penalty against perceived AI usage is due to a higher perceived likelihood of review "fake"-ness versus a more general anti-AI bias in quality evaluation. We addressed this by replicating Study 1

and further manipulating the informed agent: participants were told that each review was either generated by a human, generated by AI, or written by a human and then edited by AI; and (for each of the above) were either told nothing beyond that, or told that the review was based on a "verified in-person experience". These 6 conditions allow us to separate general anti-AI bias versus suspicions that the information in AI-generated reviews is simply fabricated. Results show that the penalty on AI perceptions is increasing in the degree of AI involvement (there is a significant negative penalty on "AI-edited", but smaller than the negative penalty on "AI-written"), and also that effects are attenuated to null differences for any reviews based on "verified" customer experiences, both for the "AI-edited" and "AI-written" conditions. This suggests that more AI involvement increases the penalty on perceived quality, and that this perceived quality penalty is primarily assessed due to the suspicion that such reviews do not represent actual customer experiences.

Finally, in a third pre-registered study, we test whether these negative perceptions of AI-perceived reviews can explain lower ratings in a random sample of 50 reviews from our field data, 25 of which were detected to involve AI usage and 25 of which were not. This allows us to both test the effect of detected AI usage on perceptions of quality, absent any confounds that may exist in our field data, and to test whether this effect is moderated by perceptions of AI specifically. Here we again find a significant negative effect of detected AI writing on perceived quality measures, with a large negative effect on perceived review authenticity and sincerity in particular, and moreover find that this effect is strongly driven by reviews that are detected to involve AI usage *and that are perceived as such by participants*. We also evaluate the relationship between participants' perceptions of AI usage in review writing and detected AI usage and find that detected AI usage, using the methodology that we employ to analyze our field data, significantly predicts participant perceptions of AI usage in our experimental study.[3]

Taken together, this suggests that perceptions of AI usage may impose a significant penalty on the perceived quality of online reviews among consumers, and that such

[3]The magnitude of the observed positive relationship is modest in magnitude, although still highly significant, which suggests imprecision in human perceptions of AI usage, closely in line with earlier work (Ma and Luo 2023). Such imprecision would likely attenuate the effects observed in our field evidence, suggesting that our estimates may be a conservative lower bound on the true effect.

an anti-AI bias may explain the strong negative effect that we estimate for Yelp.com and Amazon.com reviews detected to involve AI usage. While we cannot rule out the possibility that other candidate mechanisms may also contribute to this observed effect in the field, we here present novel evidence that such bias against perceived AI use is salient and that reviews that are detected/perceived to involve AI usage are already being rated as significantly lower-quality in the real world, likely because they are more often suspected to be fake. For managers of online review platforms, this suggests that perceived generative AI usage is a pressing danger to the quality of their product in the absence of an intervention to attenuate the threat.

At the same time, the pattern of our results also points towards a simple mitigation strategy that managers may choose to pursue: offering badged reviewer-validation programs, such as Yelp.com's "Elite" program or Amazon.com's "Vine Voice", appears to reduce the observed effect to undetectable levels, possibly because they assure readers that such reviews are not fake. Introducing or expanding such programs may prove an essential strategy component for user-generated content websites in this new era of automated text production. And for policymakers, to the extent that they seek to maintain the public goods of trustworthy user-generated ratings on online review websites, subsidizing the maintenance of such programs may help facilitate the broader social surplus of credible online reviews and ratings. Our field evidence suggests that such programs could serve as a strong first line of defense.

The rest of this paper is structured as follows. Section 2 details the relationship of this paper to prior literature. Section 3 presents the empirical analysis of our field evidence from Yelp.com and Amazon.com, including descriptive statistics on the adoption of generative AI, OLS and differences-in-differences estimates of the effect of detected AI usage on review quality, and evidence on the heterogeneity of this effect across reviewer badge status. Section 4 presents evidence from three pre-registered experiments that recreate this observed effect from the field and establish bias against perceived AI usage as a valid potential mechanism. Section 5 concludes.

2 Relation to Prior Literature

This paper contributes most directly to the new and burgeoning literature on the effects of generative AI on businesses and the economy at large. This includes Brynjolfsson et al. (2023), who found that generative AI has positive productivity impacts for online experiment participants, especially lower-skill ones; Capraro et al. (2024), who model how generative AI may exacerbate inequalities in the workplace; Otis et al. (2023), who find that generative AI leads to productivity improvements for high-skill entrepreneurs but productivity declines for low-skill entrepreneurs; Hui et al. (2023), who find that affected online freelancers see declines in both earnings and employment after the advent of generative AI; and most relevant to the current study, Burtch et al. (2023), who find that the advent of generative AI negatively harmed content quality on StackOverflow, principally due to user exit from the platform. Our study contributes novel evidence, in two new contexts and with rigorous empirical research designs, that perceived content quality of online reviews is penalized when the review is perceived to be produced using generative AI.

Second, this paper also contributes to the literature and discussion on trustworthy AI. Trustworthy AI encompasses multiple critical dimensions including robustness, fairness, explainability, privacy, safety, and accountability (Li et al. 2023a, Kaur et al. 2022, Kowald et al. 2024). While Thiebes et al. (2021) identify accuracy and accountability as core principles in this effort, a significant gap remains between abstract or high-level trustworthy AI principles and their implementation or perception in practice (Díaz-Rodríguez et al. 2023). Our research examines the harm caused by the perceived "fakeness" of AI-generated content—specifically, online reviews—and proposes interventions to enhance the trustworthiness of both the system and AI-generated outputs. Brundage et al. (2020) emphasize that verifiable claims and mechanisms to support accountability are essential for building user trust in AI systems, particularly through audit trails and transparency measures. Building on this foundation, our work contributes novel experimental evidence on how platform-level verification interventions can effectively mitigate trustworthiness failures in user-generated content platforms.

Third, this paper contributes to the literature on algorithm aversion, such that they prefer humans over AI in various domains. For example, in the domain of recom-

mendations, Longoni et al. (2019) found that people trust AI less than human doctors in making medical recommendations because they believe AI cannot account for individual needs and circumstances. Yeomans et al. (2019) also found that people relied less on algorithms for making joke recommendations because it is hard to make sense of the underlying recommendation process. In the domain of making forecasts and predictions, Dietvorst et al. (2015) found that people prefer the forecast of humans over algorithms after seeing an algorithm err, even though the algorithm can make better forecasts. Dargnies et al. (2024) looked at algorithm aversion in the domain of hiring decisions and found that people are less averse to algorithms when they are blinded, and transparency of algorithms does not have an effect. In the domain of creativity, Bellaiche et al. (2023) demonstrates a general negative bias against AI-created artworks. Chiarella et al. (2022) found that prior knowledge of AI authorship negatively influenced the aesthetic appreciation of abstract paintings. Horton Jr et al. (2023) also find a bias against AI art, but find that this can actually enhance perceptions of human creativity. Looking across domains, Castelo et al. (2019) found that people trust algorithms less in performing tasks that are perceived as subjective (vs. objective) or that are based on personal intuition and opinions. The general tendency of aversion and resistance to AI can be categorized into different psychological mechanisms: lack of understanding of AI, emotionlessness of AI, inflexibility of AI, lack of autonomy due to usage of AI, and lack of humanness in AI (De Freitas et al. 2023). This paper contributes novel findings on an analogous bias against AI usage in online reviews, providing new evidence that a similar anti-AI bias may extend to this domain and pose a present threat to online review platforms, businesses, and users.

Finally, this paper also contributes to the extensive marketing literature on eWOM creation and the determinants of credibility in online reviews. For a recent review of the eWOM creation literature, see Babić Rosario et al. (2020). More specifically, recent work on determinants of credibility on online includes Clare et al. (2016), who find that review helpfulness and review credibility are strongly linked; Craciun and Moore (2019), who find that emotional content may reduce perceived credibility of reviews, especially for female reviewers; Chevalier and Mayzlin (2006); and Lim and Van Der Heide (2015), who examine how reviewer attributes affect attitudes towards reviews through perceptions of competence. This study contributes novel evidence on

how the perceived adoption of a brand-new tool, generative AI chatbots, may further impact perceptions of authenticity and helpfulness in online reviews. Building on this, the present work also relates to the subliterature on fake reviews, including Dong et al. (2019) and He et al. (2022), among others; see Wu et al. (2020) for a more extensive literature review. This paper contributes new evidence that ChatGPT threatens to erode the trust environment around online reviews even more than before.

3 Field Evidence from Yelp.com and Amazon.com

3.1 Data

We collected consumer reviews written in both the pre- and the post-ChatGPT release periods from two popular online platforms, Yelp.com and Amazon.com, through scraping the front-facing webpages of both sites.

Yelp.com: Restaurants in San Francisco

Yelp.com is a popular website that allows users both to post reviews for brick-and-mortar local businesses and to vote on how useful, funny, and cool the posted reviews are. For this study, we gathered all Yelp.com reviews of restaurants in San Francisco, including the full text of reviews and all user-generated responses to these reviews, with the set of local businesses gathered directly from the Yelp API. We focus on San Francisco because this is a tech-savvy area that would likely feature some users who would employ ChatGPT in review production. We restrict our attention to the 78.5% of this sample that are restaurants, for a final sample of 3082 restaurants, to ensure comparability across stores. Data were collected during September 2023, and contain all reviews dating from January 2022 to August 2023. Reviews data contain review text, number of user votes on review "useful"-ness, "cool"-ness, and "funny"-ness, as well as reviewer first name, last initial, and place.[4] The data also contain markers for

[4]For example, Aaliyah S., Foster City, CA. Note that this identifying information is only approximate; while we use "reviewer" throughout this text to refer to reviews written with the same identifying information, such language refers to e.g. all reviews written by the sample of all Aaliyah S.'s from Foster City, CA. In all of our differences-in-differences specification, we present pre-period estimates to test that the observed relationships are not driven by any spurious correlation between outcomes and measurement error for reviewer designations.

whether the reviewer belongs to the "Yelp Elite Squad" group, a subgroup of users that undergo an annual vetting process from Yelp.com staff, based on review quantity and quality, in exchange for the public badge and some benefits. Restaurant data also contain location information and, for most restaurants, a coarse measure of the price range, reported on a scale of "$" to "$$$$".

Amazon.com: Camera, Photo and Video category

Amazon.com is a major online retailer that allows users to post reviews and ratings on product pages. For this study, we gathered reviews from all posted Amazon.com products that were listed under the "Camera, Photo and Video" category for the price ranges of "< $25", "$25-$50", "$50-$100", "$100-$200", and "$200+". Data were collected in October 2023, and although they include some reviews dating all the way back to the late 1990s, 90% of the sample is from 2018 to 2023. Due to practical constraints on web-scraping on Amazon, we collect the first page of reviews from each product, up to a maximum of 10 reviews per product.[5] We focused on the "Camera, Photo and Video" category to keep data collection times feasible, and because we expect that interest in technology products may correlate with likelihood to adopt GPT in review production. Review data include review text, number of user votes on review "helpful"-ness, as well as the reviewer's Amazon account page URL, which we use to uniquely identify reviewers. Data also include "Vine Voice" badges, which designate whether a reviewer belongs to the "Vine Voice" program. Similar to the Yelp "Elite" program, this is an invitation-only program for reviewers writing many high-quality reviews.[6]

Detection of GPT use in review writing

We detect the use of generative AI in the writing of the Yelp.com reviews using ZeroGPT, an online service for detecting the presence of generative AI-generated

[5] As we discuss in greater detail below, one may be reasonably concerned that our results may be meaningfully affected by ranking algorithms, which determine which reviews are placed more prominently and, in the case of Amazon, which reviews may be observed in our data. While we argue that our field results are unlikely to be solely driven by such bias, this and similar concerns motivates our extension to experiments in later sections, where we show that results qualitatively hold in out-of-sample experimental replication without any ranking confounds.

[6] For more details, see https://www.amazon.com/vine/about.

language in text. This service categorizes Yelp.com reviews according to a set of classifications, graduating from "Your Text is Human-Written" to "Your Text is Written by AI/GPT". We collect all labels delineating a high likelihood of AI/GPT use into a single binary indicator for use of AI/GPT. The ZeroGPT website claims to have a 1% to 2% false positive rate, a rate that correlates closely to the degree of AI writing detected in our samples prior to the release of ChatGPT on November 30, 2022.

However, such scorers are still by nature imperfect, and this reported false positive rate may overstate the accuracy of the ZeroGPT scorer (Ma and Luo 2023). Therefore, we refer to all reviews detected as AI-written by ZeroGPT conservatively as "detected" AI reviews, to emphasize that this may not correlate with 100% accuracy with actual AI-written reviews. In one of our pre-registered experiments, presented in Section 4, we show that participant perceptions of AI usage in review writing is significantly predicted by such ZeroGPT "detected" AI usage. Following this, we interpret and present our results of "detected" AI usage from ZeroGPT scoring as measuring the effect of *perceived* AI usage on quality metrics in the field, which may combine any effect of actual usage (to the extent that actual usage corresponds to perceived/detected usage) and any separate effect of the perception of AI usage specifically.[7]

Finally, to further account for the possibility that false positives may be associated with review attributes, we separately estimate the effect of detected AI writing both before and after ChatGPT's release in order to test for any confounds in the pre-release period. This measurement of pre-period effects helps us gauge the estimated relationship between false positives and our associated outcomes, providing a baseline for comparison for the post-GPT release measured effects.

After dropping the 23.3% reviews that do not contain enough text for ZeroGPT classification, our Yelp.com data contain approximately 29 reviews, on average, for each restaurant in our sample, for a baseline Yelp.com dataset of 79,233 scored reviews. Our Amazon.com data contain 124,513 reviews with enough text for ZeroGPT classification. For outcome measures, we focus on the perceived review quality, measured

[7]While it is outside of the scope of the current study to separately identify the effects of actual versus perceived usage in our field data, we show in our later lab studies that the effects of perceptions of AI usage on perceived quality may alone be sufficient to explain the patterns we find in our field evidence.

in terms of the number of "useful" votes or "helpful" votes that a review received from other Yelp.com users or Amazon.com users, respectively. We also present robustness checks on alternative measures that are voted on for Yelp.com, namely the number of "cool" and "funny" votes that a review received. Using review text and associated metadata, we construct a number of review characteristics for inclusion as controls, including sentiment valence using the VADER lexicon (Hutto and Gilbert 2014), review wordcount, star rating of reviews, dummies for price category and, in the case of Yelp, dummies for restaurant category. Summary statistics are presented in Table 1.[8] Finally, we remark that both Amazon.com and Yelp.com implement ranking algorithms for determining which reviews to feature most prominently; this further motivates our inclusion, in every specification, of pre-period detected-AI measures, which we suggest would capture if reviews with similar attributes are simply more likely to be ranked either higher or lower.[9]

3.2 Observed Patterns of Detected AI Adoption

We first present observational data on the patterns of adoption of generative AI in our two samples, to show how AI adoption is unfolding over time and across review types in our data.

Over Time

For both samples, we detect a statistically significant increase in detected generative AI adoption in review production after the public release of ChatGPT[10] that is also visually salient in the binscatters of detected AI usage rates across 2022 and the first 9 months of 2023, as presented in Figure 1. In the case of Yelp, AI adoption spikes in

[8] Note that, for ease of comparison across these two different contexts, we standardize our outcome variables to express effects in terms of standard deviation shifts in perceived quality in all regressions.

[9] Nonetheless, one might expect that this could lead to an "acceleration" of observed field effects if upvotes tend to beget more upvotes, as reviews with a small mass of upvotes may earn more upvotes by virtue of their more competitive placement. We argue that such a bias would be unlikely to qualitatively change the direction of effects. More broadly, this and similar concerns around potential biases in field data motivate our focus on experimental results in later sections, which are ensured to be free of such biases.

[10] Regressing an indicator for the post-release period on a dummy for whether AI was detected in the writing of the review, for each respective sample we find a highly significant increase: $\beta_{Amazon} = 0.019$, $t_{124971} = 15.9$, $p < 0.001$, and $\beta_{Yelp} = 0.007$, $t_{79231} = 6.88$, $p < 0.001$.

Table 1: Summary Statistics

	Yelp.com Sample		Amazon.com Sample	
	Mean	SD	Mean	SD
"Useful" votes	1.33	(5.51)	.	(.)
"Funny" votes	0.49	(3.26)	.	(.)
"Cool" votes	1.05	(4.97)	.	(.)
"Helpful" votes	.	(.)	6.52	(32.61)
Star rating	4.15	(1.21)	4.29	(1.06)
Valence	0.22	(0.11)	0.68	(0.49)
Word count	108.50	(88.13)	166.83	(193.81)
Observations	79233		124513	

Notes: Summary statistics of outcome measure and controls. Sample based on Yelp reviews for restaurants in San Francisco area and Amazon reviews for "Camera, Photo and Video" category, with GPT use detected using ZeroGPT. Valence measured using the VADER lexicon.

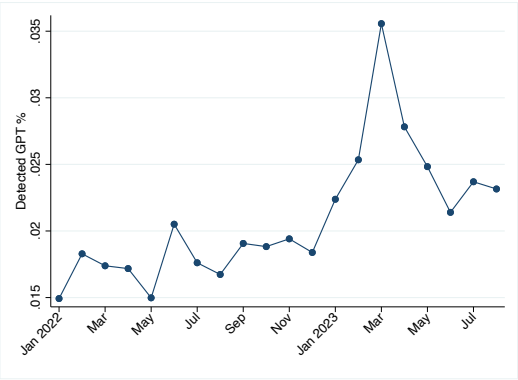
the first few months of 2023 from around 1.5% to 3.5%, before moderating over the summer to 2.5%; while in the case of Amazon, detected AI use doubles from around 3% to approximately 6% after the release of ChatGPT.

We remark that while the scale of the y-axis is wide due to the spike in March of 2023, the final level of detected GPT use in July 2023 is still approximately 50% higher than July 2022 on Yelp.com, with a level of around 2.3% versus 1.5% previously. This is a similar order of magnitude to the rough doubling of detected GPT use on Amazon.com over the same period, from 3% to 6%.

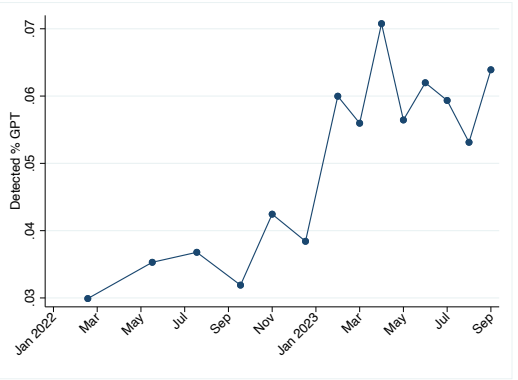
In both cases, we see that there is a meaningful proportion of reviews detected as generated with AI prior to the release of ChatGPT. This suggests that around 2%-3% of detected AI-involved reviews arise false positives from the ZeroGPT detection

Figure 1: Adoption of Generative AI Over Time

(a) Yelp.com



(b) Amazon.com



Notes: Adoption of generative AI over time for Yelp restaurant reviews in San Francisco and Amazon reviews of "Camera, Photo and Video" category products. Y-axis shows the detected percentage of reviews that were generated using generative AI, X-axis shows the date. Use of generative AI scored using ZeroGPT.com. Sample based on 79,233 reviews gathered from Yelp.com for the universe of restaurants in San Francisco, and 124,513 reviews gathered from Amazon.com for the "Camera, Photo and Video" category.

software, roughly in line with the reported rates from ZeroGPT.[11] In light of this, in order to ensure that our results are not driven by false positive rates in GPT detection, in our empirical analysis we separately analyze the effects of detected GPT use prior to the broad release of ChatGPT versus afterwards, and focus primarily on the difference between the pre-release estimates and post-release estimates as the measure of detected ChatGPT adoption's effect on review production. This measured pre-period effect should capture the association between the types of reviews likely to be false-positively detected as using AI in writing, and thus helps ground expectations as to the direction and magnitude of bias from the inclusion of some proportion of false positives in the post-GPT release period.

[11]It may also be the case that some reviewers were using earlier versions of LLMs to produce reviews; we cannot separately identify them here.

Across Review Characteristics

Next, we inspect adoption over different types of reviews. Namely, we inspect whether generative AI is more likely to be employed in the production of a 5-star review, or a 1-star review; a positive or a negative review; or a longer or shorter review (as defined by wordcount).

Broadly speaking, reviews written with generative AI versus reviews written without appear to show similar J-distributions of star ratings, presented in Figure 2. This suggests that GPT-written reviews are not exclusively being used to "flood" either 1-star or 5-star reviews, at least in our data samples, but are instead being adopted by users with a similar spectrum of feedback and ratings as in the part of the sample detected to be written without AI assistance. Distributions of review valence and review wordcount across GPT-written and human-produced reviews are also highly similar, presented in Appendix Figures A4 and A5. The only salient difference is that GPT-written reviews appear to be more scarce in the lowest wordcount buckets, which is consistent with the common belief that such reviews are less likely to be very short.

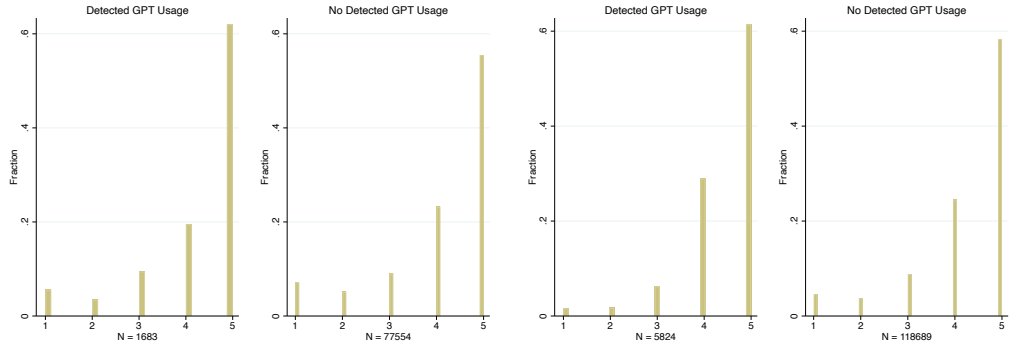
3.3 Econometric Specification

To assess the impact of generative AI on perceived quality and volume of reviews, we employ three primary econometric specifications. First, we use a simple ordinary-least-squares (OLS) regression on a dataset of scored individual reviews to investigate the association of generative AI use and observed perceived review quality, as measured by the number of user-generated "useful" or "helpful" votes on the given review q_{ijt}, where GPT_{ijt} is a binary measure of whether generative AI was detected in the production of review i by reviewer j in period t and $\mathbf{X_{ijt}}$ represents a vector of controls. $\beta_{GPT-pre}$ measures the effect of detected AI usage on perceived review quality in the pre-ChatGPT-release period of January 2022 through November 2022, and $\beta_{GPT-post}$ measures the effect of detected AI usage on perceived review quality in the post-ChatGPT-release period of December 2022 to September 2023.

$$q_{ijt} = GPT_{ij,t=0}\beta_{GPT-pre} + GPT_{ij,t=1}\beta_{GPT-post} + \mathbf{X_{ijt}}\beta_{\mathbf{X}} + \epsilon_{ijt} \tag{1}$$

Figure 2: Star Rating Distribution:
Human versus Detected AI

(a) Yelp.com (b) Amazon.com



Notes: Star rating distribution of reviews, stratified by detected AI usage. Y-axis of each panel shows the fraction of reviews at each star rating, while X-axis shows the associated star rating. Use of generative AI scored using ZeroGPT.com. Sample based on 79,233 reviews gathered from Yelp.com for the universe of restaurants in San Francisco, and 124,513 reviews gathered from Amazon.com for the "Camera, Photo and Video" category.

With this specification, we capture the effect of detected AI usage on perceived review quality on average. In particular, the differences between $\beta_{GPT-post}$ and $\beta_{GPT-pre}$ allow us identify the observed change in the effect of the novel use of ChatGPT in Yelp.com or Amazon.com review production, respectively.

We use a similar specification to estimate the effect of generative AI on log quantity of reviews written per reviewer, N_{jt}, after collapsing the data to the reviewer j-by-period t level for the periods of January 2022 to November 2022 and December 2022 to September 2023 or August 2023 for Amazon.com or Yelp.com, respectively.[12] We then estimate the effects on quantity with separate coefficients for pre- and post-release effects of detected AI use in writing, as before:

$$log(N_{jt}) = G\bar{P}T_{j,t=0}\beta_{GPT-pre} + G\bar{P}T_{j,t=1}\beta_{GPT-post} + \mathbf{X_{jt}}\beta_{\mathbf{X}} + \epsilon_{jt} \tag{2}$$

Finally, to investigate results using the most extensive set of controls available to

[12]We scale N_{jt} for the post-period by $\frac{11}{9}$ or $\frac{11}{10}$ for Yelp.com and Amazon.com, respectively, to standardize the N to account for slightly differing period lengths.

us, we further implement a specification that estimates the effect of detected AI usage prior to ChatGPT's release and after in a two-period event-study design, to control for possible differing selection into generative AI use:

$$\bar{q}_{jt} = G\bar{P}T_{j,t=0}\beta_{GPT-pre} + G\bar{P}T_{j,t=1}\beta_{GPT-post} + \mathbf{X_{jt}}\beta_{\mathbf{X}} + 1_{j=J}\delta_j + 1_{t=T}\gamma_t + \epsilon_{jt} \quad (3)$$

In this specification, the estimated $\beta_{GPT-post}$ is identified from the effect of detected AI usage after ChatGPT was released, controlling for period and reviewer fixed effects. This allows us to identify the effect of generative AI use on the usefulness of reviews of the same identified reviewer, pre- versus post-release.[13] In the appendix, we also present results using our baseline OLS sample, but with a simple author fixed effect includes; results are qualitatively highly similar between this specification and our baseline two-period differences-in-differences specification.

Control Inclusion

In the context of analyzing the effect of ChatGPT on perceived quality, one may be concerned that including controls for specific review characteristics (e.g. wordcount) may expose our specification to concerns that such characteristics are "bad controls", since these characteristics may lie on the causal pathway by which detected ChatGPT use affects perceived review quality. To account for this concern, in all of our tables we include specifications with no controls included (i.e., with the $\mathbf{X}$ term excluded from the above equations), as well as specifications with a full set of rich controls included. As shown below, we broadly find similar results across such "no control" and "control" specifications, suggesting that concerns about bad controls are unlikely to be salient for our specifications.

[13]Note that this specification, since it is a two-period differences-in-differences design, is not vulnerable to the issues with differences-in-differences methodology related to heterogeneous cohort effects pointed out by Callaway and Sant'Anna (2021) and others—this is not a 'staggered' differences-in-differences design, but a single two-period event study.

3.4 Empirical Results

3.4.1 Effect of Detected Generative AI Usage on Output Quality

We first examine the impact of detected generative AI use on the perceived quality of reviews, measured by review "useful"-ness votes from Yelp.com users or review helpfulness votes from Amazon.com users.[14]

Results are presented in Tables 2 and A13. We find that across the board, detected generative AI use in the post-ChatGPT release period is associated with lower-quality production, with effects ranging from 10.4% to 14.9% of a standard deviation decrease in quality from using generative AI. On both Amazon.com and Yelp.com, reviews are less useful to other users when detected as written using ChatGPT in the post-ChatGPT release period. Pre-period associations between perceived review quality and detected AI use in review writing differ across settings, with a tightly estimated zero effect for Yelp.com reviews and a strongly positive association between detected AI use and perceived review quality for Amazon.com reviews. In both cases, the estimates clearly show that the significant negative effect found in the post-period is not confounded by pre-period effects.[15]

When controlling for average valence of the review text, review word count, the star rating of the review, dummies for price range and (in the case of Yelp) dummies for restaurant category, the estimated effect in both cases remains large and significant. We also find that these negative effects are robust to using alternative measures of perceived review quality, the "cool"-ness and "funny"-ness of Yelp.com reviews, presented in Appendix Tables A14 and A15. While the effect on "useful" votes is the greatest, this suggests that part of the effect on quality may be related to the differences in the style of the writing, as measured by the response of other users voting on review "funny" and "cool" dimensions. Across all these Yelp.com specifications,

[14]Note that we standardize outcome variables to express effects in terms of standard deviation shifts in perceived review quality. We avoid conventional log(x) formulae due to the high prevalence of zero-valued observations and do not use standard log(1+x) adjustments due to econometric issues with that formula recently highlighted in Chen and Roth (2024); nonetheless, for robustness we present results using log(x) in Appendix Table A12. Results are qualitatively highly similar.

[15]In Appendix Table A10, we add writing style controls from Boyd et al. (2022) to our Amazon regression model and show that this significantly attenuates the pre-period association between "detected GPT" and review quality, suggesting that this pre-period positive association is driven by writing style, and that the writing style associated with "detected GPT" having a positive association with review quality prior to ChatGPT's release.

we find small and statistically null results for the effect of detected generative AI use in the pre-period, as in our baseline specification.

Additional Validity Check: Post-Period Detected AI Usage and Pre-Period Differences

While we interpret the null coefficients on pre-period detected AI usage as evidence against the idea that pre-trends may explain the negative effects we observe in the post-period, one may remain concerned that detected AI usage in either period measures false positives in the pre-period versus actual usage in the post-period—and while it helps to rule out the possibility that false positives drive our results, it does not rule out the possibility that those who later adopt AI simply write worse reviews in both periods. We investigate this directly with an author-level regression of pre-period helpfulness/usefulness differences between authors detected to use AI in the post-period and authors not detected to use AI in the post-period. Results are presented in Appendix Table A7. We find null pre-period review quality differences between authors detected to use AI in the post-period and others, both with and without the inclusion of controls, and moreover point estimates are insignificantly positive. From this, we infer that it is highly unlikely for innate differences across adopting and non-adopting review authors drives our results.

Heterogeneity by Reviewer Badge Presence

Given that our results are measuring the effect of detected/perceived AI usage, one may expect that the presence of a clear indicator of verified "expert" involvement in the review creation process may be a meaningful moderator of the effect, impacting either the extent to which similar writing is perceived to rely on low-effort or "spam" AI usage or offering a countervailing indicator of quality that may attenuate any perception-based bias.

To explore this potential effect heterogeneity, we break down our sample according to Yelp.com's designation of "Elite" versus non-"Elite" reviewers, and Amazon.com's designation of "Vine Voice" versus non-"Vine Voice" reviewers. (For convenience, we refer to either "Elite" or "Vine Voice" reviewers as "badged" reviewers below, and reviewers that do not belong to either category as "non-badged" reviewers.) In each

Table 2: Effect of Detected AI Use on Yelp.com Review "Useful" Votes

$\beta_{GPT-pre}$	-0.011	-0.012
	(0.038)	(0.036)
$\beta_{GPT-post}$	-0.104**	-0.165***
	(0.032)	(0.031)
Controls		X
N	79233	79233

Table 3: Effect of Detected AI Use on Amazon.com Review "Helpful" Votes

$\beta_{GPT-pre}$	0.165***	0.080***
	(0.019)	(0.018)
$\beta_{GPT-post}$	-0.149***	-0.142***
	(0.018)	(0.017)
Controls		X
N	124513	124513

Notes: Effect of adoption of ChatGPT in review production on review quality. Sample based on Yelp reviews for restaurants in San Francisco area and reviews for products in "Camera, Photo and Video" category of Amazon, with GPT use detected using ZeroGPT. Controls include star rating of review, wordcount of review, average valence of review, price level dummies, and for Yelp restaurants, restaurant type dummies. Outcome variables standardized to express effects in terms of standard deviation shifts in quality. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

case, these are 'earned' status badges for reviewers, acquired in recognition of high reviewer skill and high output, and contingent on continued high-value production in terms of useful/helpful reviews (for more information, see https://www.yelp.com/elite and https://www.amazon.com/vine/about). While we rely on these badges to divide our sample into two separate groups and structure the heterogeneity of the effect, we are unable to separate in this context whether any observed differences between them is due to the effect of the badge itself or due to differences in worker type.[16] Since we are better able to investigate the potential mechanism of consumer perceptions of AI usage within the scope of this study, we highlight that candidate explanation throughout, but we note here that we cannot rule out that other mechanisms associated with other correlates of badge/non-badge status may also play an important role.

Results are presented in Appendix Tables A16 and A17. Using the same specification as above, stratified by reviewer type, we find that our measured effect on quality manifests for both badged and non-badged reviewers: in the case of Yelp.com, we find a slightly stronger negative association for badged reviewers in terms of point estimates, while in the case of Amazon.com, we find a slightly stronger negative association for non-badged reviewers. In all cases, we find a negative association between post-ChatGPT release generative AI usage and review reported quality that is not confounded by any pre-period negative effects.

Differences-in-Differences Estimates

While this evidence is strongly suggestive of a negative relationship between detected AI usage in the review-writing process and perceived review quality, it is possible that this simple specification is confounded by differential selection into the use of generative AI in review writing (assuming an association between actual use and detected use). If worse reviewers are more likely to adopt generative AI, lower-quality reviews could be associated with detected generative AI use outside of a causal relationship between detected generative AI and output quality. Moreover, the degree of this selection effect may differ across groups, similarly confounding our comparisons

[16] In section 4, we present experimental evidence that communicating the human-ness of the writer can attenuate any AI penalty associated with perceptions of AI, even for writing generated by AI.

across badged and non-badged groups. Separately, one may also be concerned that more recent reviews are both more likely to feature GPT usage and have had less time to accumulate helpfulness or usefulness upvotes, leading to a spurious negative relationship between GPT usage and quality ratings.

To control for these and other possible confounders, we collapse our data into reviewer-level averages for the two periods of January 2022 to November 2022 and December 2022 to September 2023, in order to assess the effect of generative AI writing on a given reviewer using a simple two-period differences-in-differences empirical design. This allows us to inspect the changes in perceived review quality within a given reviewer as they adopt generative AI into their review production process across periods. Then if we find an effect of detected generative AI to be significantly different from the estimated effect for the pre-period, this implies that reviewers whose detected use of generative AI increased from period-to-period saw a concomitant change in the observed quality of their output, and we can feel more assured that this effect is driven by proportional changes within an identified reviewer rather than spurious selection effects or other reviewer-specific confounds, and with period fixed effects as well to control for time-related confounds.

This strategy is most viable for our Yelp.com sample, where we are able to match nearly 5,000 reviewers across periods. For our Amazon.com data, we are only able to match under 1,300 reviewers across periods. We present results for both, but with the caveat that Amazon.com differences-in-differences estimates are underpowered relative to our Yelp.com estimates. Note that in this regression, our AI usage variable is now a reviewer-by-period average and not a binary review-level indicator; to ease comprehension, we standardize the AI adoption variables so that results are presented in terms of the effect of a one-standard-deviation increase in AI usage in review production in each respective sample.

Results of this differences-in-differences regression on standardized perceived review quality are presented in Tables 4 and 5. For the Yelp.com context, we uncover a strongly significant negative effect of detected ChatGPT adoption on perceived review quality, even when controlling for reviewer and period fixed effects: a standard deviation increase in reviewer AI usage in the post-ChatGPT release period leads to a 6.5% standard deviation decrease in review "useful"-ness. For the Amazon.com

Table 4: Effect of Detected AI Use on Yelp.com Review "Useful" Votes: Differences-in-Differences

$\beta_{GPT-pre}$	-0.007	-0.007
	(0.010)	(0.010)
$\beta_{GPT-post}$	-0.063***	-0.065***
	(0.010)	(0.010)
δ_j	X	X
γ_t	X	X
Controls		X
N	9682	9682

Table 5: Effect of Detected AI Use on Amazon.com Review "Helpful" Votes: Differences-in-Differences

$\beta_{GPT-pre}$	-0.028	-0.029
	(0.016)	(0.016)
$\beta_{GPT-post}$	-0.024	-0.033
	(0.029)	(0.029)
δ_j	X	X
γ_t	X	X
Controls		X
N	2578	2578

Notes: Difference-in-differences estimates of the effect of adoption of ChatGPT in review production on review quality. Sample based on Yelp reviews for restaurants in San Francisco area and reviews for products in "Camera, Photo and Video" category of Amazon, with GPT use detected using ZeroGPT. Controls include star rating of review, wordcount of review, average valence of review, price level dummies, and for Yelp restaurants, restaurant type dummies. Outcome variables and covariates standardized to express effects in terms of standard deviation shifts. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

context, with our smaller sample size we do not have enough precision to estimate statistically significant effects, but the point estimates suggest that a standard deviation increase in reviewer AI usage in the post-period leads to a 3.3% standard deviation decrease in review "helpful"-ness. 95% confidence intervals for the Amazon.com sample estimates are inclusive of the estimates from the Yelp.com sample as well.

Finally, we extend this differences-in-differences regression to stratified samples restricted to either badged and non-badged reviewers, as in section 4.1.1, to estimate how this effect differs across reviewer types. Results are presented in Tables 6 and 7. Here we see a clear distinction in the effects of detected AI usage across reviewer types: the negative relationship with quality is consistently much larger for non-badged reviewers. For Yelp.com reviews, the overall observed effect appears to be entirely driven by non-badged reviewers, with a precisely estimated zero effect for "Elite" reviewers in this differences-in-differences specification versus a highly significant 17.2% decline in "helpful" votes for reviews by non-"Elite" reviewers detected to use generative AI in the post-release period. For the much smaller sample of Amazon.com reviews, again both point estimates are insignificant but suggest that the negative effect for non-badged reviewers is twice as large, comparing the point estimates of -0.029 and -0.054 between "Vine Voice" and non-"Vine Voice" reviewers.

Notably, the distinction between the effects across these groups is clearer and more consistent in these tables than in the observational specification of Appendix Tables A16 and A17. This suggests that not only does the effect differ across these groups, but also selection patterns may differ as well: the OLS estimates of the effect of detected AI usage on "Elite" Yelp.com reviewers may plausibly be driven by negative selection into ChatGPT use, where the least-skilled "Elite" reviewers select into using generative AI in review production, while baseline effects on non-"Elite" Yelp.com reviewers may plausibly be attenuated by *positive* selection into ChatGPT use, where the most-skilled non-badged reviewers choose to adopt generative AI in their review production. Results for Amazon.com are too noisy to make inferences on the plausible direction of selection effects, if any.

Overall, it appears that perceived AI usage is broadly associated with a major penalty on perceptions of review quality, and with the effect driven by non-badged

Table 6: Effect of Detected AI Use on
Yelp.com Review "Useful" Votes:
Differences-in-Differences

	"Elite"		Non-"Elite"	
$\beta_{GPT-pre}$	-0.007	-0.006	-0.005	-0.006
	(0.009)	(0.009)	(0.023)	(0.023)
$\beta_{GPT-post}$	0.006	0.005	-0.168***	-0.172***
	(0.009)	(0.009)	(0.023)	(0.023)
δ_j	X	X	X	X
γ_t	X	X	X	X
Controls		X		X
N	4290	4290	5118	5118

Table 7: Effect of Detected AI Use on
Amazon.com Review "Helpful" Votes:
Differences-in-Differences

	"Vine Voice"		Non-"Vine Voice"	
$\beta_{GPT-pre}$	-0.013	-0.015	-0.045	-0.040
	(0.012)	(0.012)	(0.031)	(0.031)
$\beta_{GPT-post}$	-0.024	-0.029	-0.033	-0.054
	(0.023)	(0.023)	(0.050)	(0.049)
δ_j	X	X	X	X
γ_t	X	X	X	X
Controls		X		X
N	968	968	1560	1560

Notes: Difference-in-differences estimates of the effect of adoption of ChatGPT in review production on review quality, stratified by reviewer category. Sample based on Yelp reviews for restaurants in San Francisco area and reviews for products in "Camera, Photo and Video" category of Amazon, with GPT use detected using ZeroGPT. Controls include star rating of review, wordcount of review, average valence of review, price level dummies, and for Yelp restaurants, restaurant type dummies. Outcome variables and covariates standardized to express effects in terms of standard deviation shifts. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

reviews. This pattern could be consistent with an effect of actual generative AI usage where only low-skill workers use the technology to produce worse output, or with an effect of perceived generative AI usage that is driven entirely by non-badged reviews. The first hypothesis would be inconsistent with some prior evidence showing an especially beneficial effect of generative AI usage on the productivity of low-skill workers (Brynjolfsson et al. 2023), but the second hypothesis would be consistent with prior evidence showing a negative perceived-quality penalty for creative art known to be produced by visual generative AI (Bellaiche et al. 2023, Chung 2023).

3.4.2 Effect of Generative AI on Output Quantity

Finally, we examine the effects of detected AI usage on observed quantity of reviewer output, using the specification described in equation 2. To perform this analysis, we collapse our data into a reviewer-by-period dataset for the pre-period of January 2022 to November 2022 and the post-period December 2022 to September 2023 or August 2023 for Amazon.com and Yelp.com, respectively, and count how many reviews each reviewer produces in either period. We then regress the log number of reviews on the averages of AI-usage indicators for each reviewer X period observation, for both the pre- and post-ChatGPT release period, analogous to our specification for quality. Full analyses are presented in Appendix Section A1; we detect a modest post-period increase in output quantity using OLS, but null effects in differences-in-differences regressions.

4 Experimental Evidence

To deepen our evidence, we turn to three pre-registered online experiments to help trace out the causal path between detected AI usage and review quality ratings. In the first experiment, we manipulate both perceptions of AI-generated writing and actual presence of AI-generated writing to isolate which aspect drives the effect. In the second experiment, we re-run the first experiment but clarify to all participants that reviews are based on verified customer experiences, to separate out any part of the effect arising from perceptions of fake versus real reviews. In the third experiment, we present reviews from our field data, 25 detected human-written and 25 detected

AI-written, to online survey participants and solicit both their perceptions of review quality and their perceptions of whether it was majority-AI-written. In all studies, our aim was to examine the extent to which the effect could be explained by negative quality effects of actual usage versus negative quality effects of perceived AI usage.[17]

With this follow-up investigation, we seek to decompose the pattern of our field evidence, where we found a large effect in differences-in-differences estimation that was driven by non-badged reviews. While this pattern regarding badges is suggestive that perceptions may play a role, we highlight that these observed field results may be driven by some combination of effects of actual AI usage and effects of perceived AI usage, including the possibility that they are entirely driven by effects of actual usage or entirely driven by effects of perceptions of usage. We cannot separate these effects in our field data, where we do not separately observe perceptions versus actual usage, and indeed presume that perceptions and actual usage are likely very highly correlated. Similarly, we cannot test in our field data whether our AI detection tool corresponds to perceptions, and so turn to lab experiments to help fill this gap.

We begin with a simple experiment to examine whether reviews that are actually AI-written, versus reviews that are explicitly described as AI-written, are more likely to drive the observed negative effect we find in the field.

4.1 Experiment 1: Consumer Perception of Human vs. AI-Written Reviews

In Experiment 1, we examined how consumers perceive reviews generated by either AI or humans.[18] We also manipulated the informed agent by telling participants the review was generated by AI or human.

Method

We recruited 292 participants via Prolific (49.6% female, $\mu_{age} = 36.0$) who passed our attention check question. Participants were randomly assigned to one of four

[17] Note that we use "bias" here not to mean that such bias is taste-based and not based on rational expectations. We use this term only to mean that there is a penalty applied to perceptions of quality when AI usage is inferred.

[18] Pre-analysis plan for this study may be accessed at https://researchbox.org/3239&PEER_REVIEW_passcode=WHBXDN.

conditions in a 2 (writing agent: AI vs. human) x 2 (informed agent: AI vs. human) between-subjects design.

Participants were told they would read two reviews of products listed on Yelp.com and be asked to evaluate the quality of the reviews. The reviews with humans as the writing agents were 25 reviews randomly selected from our field data online review dataset of Yelp.com reviews. After selecting the human-written reviews, we asked ChatGPT to generate an AI-written version for each review based on the following prompt: "Please rewrite the following survey in your voice." The 25 ChatGPT-generated reviews served as the AI-written reviews.

Participants were randomly presented with two reviews based on the writing agent condition they were randomly assigned to (AI vs. human). Those in the AI-agent condition read reviews generated by ChatGPT, while those in the human-agent condition read reviews generated by humans. We also presented the product names, product photos, and star ratings with the review text. Participants were also informed of the review writer based on the informed agent condition they were assigned to (AI vs. human). Those in the informed AI condition read: "Please read the following review that is written by ChatGPT." Those in the informed human condition read: "Please read the following review that is written by the customer herself."

After being informed of the writing agent and reading the review, participants were asked to rate the reviews on the following dimensions: authenticity, helpfulness, usefulness, persuasiveness, sincerity, convinced by the review, and willingness to purchase the products on a seven-point Likert scale (1 = not at all, 7 = very much). The participants rated the two reviews separately. Lastly, participants answered an attention check question and demographic questions.

Results

Results from experiment 1 are presented in Table 8. Impressions of reviews split strongly along the dimension of perceptions of AI usage, as controlled through the express communication in the stimuli. Compared to the excluded category of stimuli that was communicated to be human and was detected to be non-AI, both treatments that communicated AI usage in the subsequent writing were rated as lower-quality across our full set of dimensions, with an especially large penalty for measures of

Table 8: Effect of Actual and Communicated AI Use on Review Quality Perceptions

	Authentic	Helpful	Useful	Persuasive	Sincere
$\beta_{CommunicatedAI \times ActuallyAI}$	-1.769***	-0.554**	-0.640**	−0.396†	-1.330***
	(0.200)	(0.190)	(0.200)	(0.211)	(0.202)
$\beta_{CommunicatedAI \times ActuallyHuman}$	-2.529***	-1.276***	-1.536***	-1.064***	-2.157***
	(0.198)	(0.188)	(0.197)	(0.209)	(0.200)
$\beta_{CommunicatedHuman \times ActuallyAI}$	-0.129	0.265	0.143	0.257	-0.094
	(0.195)	(0.186)	(0.195)	(0.206)	(0.197)
$\beta_{CommunicatedHuman \times ActuallyHuman}$	.	.	.	.	.
(Excluded Category)	(.)	(.)	(.)	(.)	(.)
N	584	584	584	584	584

Notes: Effect of different combinations of actual and communicated AI usage on review quality perceptions. Sample drawn from Prolific and stimuli drawn from field data (Yelp.com) reviews that were detected to involve no GPT use, with GPT use detected using ZeroGPT. For actual AI usage, we then asked ChatGPT to rewrite said baseline no-detected-AI reviews "in [its] own voice". Usage was communicated explicitly to be either "written by the customer herself" or "written by ChatGPT". † = 10% significance, * = 5% significance, ** = 1% significance, *** = 0.1% significance.

authenticity and sincerity. We emphasize that this highly significant and broad-based quality differential exists between participants viewing identical stimuli, with only different framing, either expressed as human-written or ChatGPT-written. We further find no significant difference between quality perceptions of reviews that were actually written using generative AI and those that were detected as non-AI so long as both were communicated to be human-written. This evidence supports the hypothesis that impressions of AI usage may drive the pattern of results observed in our field data so long as perceptions of AI usage among consumers tracks with detected AI usage as labeled with ZeroGPT.

We remark here as well that this pattern of effects accords well with our observed effects in the field. In our field data, reviews are not labeled as AI-written or human-written, and so consumers never encounter AI-written reviews that are explicitly described as human-written (except, one might argue, in the case of "badged" reviews). Instead, consumers have to infer the likelihood that a review is AI-generated from the

text, leading to a strong entanglement between actual AI usage and perceptions of AI usage. As such, our effects showing that perceptions of AI usage appear to solely drive negative quality perceptions are consistent with the patterns of our field data, and moreover are consistent with the pattern of significant results for non-badged reviews but null results for badged reviews. We examine this distinction more closely in the following study.

4.2 Experiment 2: Consumer Perception of *Verified* Human vs. AI-Involved Reviews

In Experiment 2, we aim to accomplish two goals. First, we further explored whether the above documented negative penalty for perceptions of AI usage arises due to perceptions of review "fakeness", i.e., the likelihood that the review is not based on an actual experience, or simply due to general AI aversion. To address this question, we manipulated the informed agent by telling the participants that the review was generated by a human, AI, or written by a human then edited by AI. We hypothesized that differences in effect magnitude off AI penalty would reflect the component of suspicions of fakeness in addition to AI aversion in general. Second, we examine whether stating reviews as "based on a verified customer experience" serves as an effective intervention to combat the perception of review fakeness.[19] The findings provide insight into whether verified badges provided by platforms can effectively increase consumer trust in reviews.

Method

We recruited 587 participants via Prolific (49.7% female, $\mu_{age} = 43.7$) who passed our attention check question. Participants were randomly assigned to one of six conditions in a 3 (informed agent: AI vs. AI-edit vs. human) x 2 (verified experience: yes vs. no) between-subjects design.

Participants were told they would read two reviews of products listed on Yelp.com and were asked to evaluate the quality of the reviews. The reviews were 25 reviews

[19] Pre-analysis plan for this study may be accessed at https://aspredicted.org/d365-srw4.pdf.

randomly selected from our field data online review dataset of Yelp.com reviews. Participants were randomly presented with two reviews. We also presented the product names, product photos, and star ratings with the review text. Participants were also informed of the review writer based on the informed agent condition they were assigned to (AI vs. AI-edit vs. human). Those in the informed AI condition read: "Please read the following review that is written by ChatGPT." Those in the informed AI-edit condition read: "Please read the following review that is written by the customer herself and edited by ChatGPT." Those in the informed human condition read: "Please read the following review that is written by the customer herself." In addition, we further told those in the verified-experience condition that the review was based on a "verified in-person experience". Those in the no-verified experience condition did not receive this information.

After being informed of the writing agent and reading the review, participants were asked to rate the reviews on the following dimensions: authenticity, helpfulness, usefulness, persuasiveness, sincerity, convinced by the review, and willingness to purchase the products on a seven-point Likert scale (1 = not at all, 7 = very much). They also indicate whether they think the review writer actually had the in-person experience by selecting "yes" or "no". The participants rated the two reviews separately. Lastly, participants answered an attention check question and demographic questions.

Results

Results from experiment 2 are presented in Table 9. As pre-registered, we reported results that excluded participants in the "verified" conditions who did not believe in the platform verification. Specifically, we excluded those in the "verified" conditions who did not believe that the review writer actually had the in-person experience. We report results that include these participants in the appendix.

Again, impressions of reviews split strongly along the dimension of perceptions of AI usage when the experience is not verified. Compared to reviews communicated as human-written, those indicating AI involvement—whether written by AI or edited by AI—were rated lower in quality across all dimensions. We also observed attenuated effects in the AI-edit condition, suggesting that the penalty for AI usage goes beyond

Table 9: Effect of Communicated AI usage and Verified Experience on Review Quality Perceptions

	Authentic	Helpful	Useful	Persuasive	Sincere
$\beta_{CommunicatedAI\ X\ Verify}$	$-0.301^{\dagger}$	0.074	0.090	-0.060	-0.113
	(0.158)	(0.167)	(0.164)	(0.177)	(0.153)
$\beta_{CommunicatedAI\ X\ NoVerify}$	-1.348***	-0.843***	-0.758***	-0.884***	-1.045***
	(0.147)	(0.156)	(0.152)	(0.165)	(0.144)
$\beta_{CommunicatedAIEdit\ X\ Verify}$	$-0.259^{\dagger}$	-0.129	-0.132	-0.250	-0.176
	(0.153)	(0.162)	(0.158)	(0.171)	(0.150)
$\beta_{CommunicatedAIEdit\ X\ NoVerify}$	-0.708***	-0.641***	-0.620***	-0.703***	-0.696***
	(0.147)	(0.156)	(0.152)	(0.165)	(0.144)
$\beta_{CommunicatedHuman\ X\ Verify}$	0.155	0.213	0.192	-0.090	$0.273^{\dagger}$
	(0.153)	(0.163)	(0.159)	(0.172)	(0.150)
$\beta_{CommunicatedHuman\ X\ NoVerify}$	.	.	.	.	.
(Excluded Category)	(.)	(.)	(.)	(.)	(.)
N	1,088	1,088	1,088	1,088	1,088

Notes: Effect of different combinations of communicated AI usage and verification of experiences on review quality perceptions. Sample drawn from Prolific and stimuli drawn from field data (Yelp.com) reviews that were detected to involve no GPT use, with GPT use detected using ZeroGPT. Usage was communicated explicitly to be either "written by the customer herself", "written by the customer herself and edited by ChatGpt", or "written by ChatGPT". Verification of experience was communicated by explicitly to be "based on a verified in-person experience".† = 10% significance, * = 5% significance, ** = 1% significance, *** = 0.1% significance.

general AI aversion and is also driven by perceived fakeness of experiences. In addition, we found that "verified" reviews profoundly attenuated the negative effects of AI usage: compared to reviews that were not verified and involved AI usage, reviews that were verified showed null effects across the review quality dimensions, even when AI usage is communicated. This supports the hypothesis that generative AI usage in online reviews is a crisis of trust in particular, not necessarily of quality. The results suggest verification of reviews as a potentially effective intervention for platforms to increase trust in reviews.

4.3 Experiment 3: Consumer Perception of Reviews from Field Data

In Experiment 3, we investigated whether these patterns hold in a random sample of reviews from our field data (without any description of the agent/writer), and further investigated whether participants could differentiate reviews written by AI vs. humans without labeling.[20] Our aim was to examine whether the quality penalty that we find in the field is also present in lab participant ratings, and whether this penalty was significantly moderated by perceptions of AI usage in particular.

Method

We recruited 286 participants via Prolific (48.6% female, $\mu_{age} = 35.9$) who passed our attention check question. Participants were told they would read six reviews of products listed on Yelp.com and be asked to evaluate the quality of the reviews. These six reviews were randomly selected from a list of 50 reviews that were selected from our field dataset of online reviews from Yelp.com. We randomly picked 25 reviews from our field data that were classified by our algorithm as AI-generated and 25 reviews that were classified by our algorithm as human-generated.

Participants were randomly presented with six reviews in a sequence. We also presented the product names, product photos, and star ratings with the review text. After reading each review, participants were asked to rate the reviews on the following dimensions: helpfulness, quality, usefulness, authenticity, persuasiveness, and

[20]Pre-analysis plan for this study may be accessed at https://researchbox.org/3238&PEER_REVIEW_passcode=XTUUGW.

Table 10: Relationship Between
Perceived and Detected AI Usage

	Perceived as 50%+ AI-written	Perceived % AI-written
$\beta_{DetectedAI}$	0.068**	6.537***
	(0.022)	(1.457)
N	1716	1716

Notes: Effect of detected AI usage on perceptions of AI usage in online survey. Sample drawn from Prolific with stimuli randomly drawn from field data (Yelp.com reviews) with GPT use detected using ZeroGPT, stratified to comprise of 25 reviews with detected AI usage and 25 reviews with no detected AI usage. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

sincerity on a seven-point Likert scale (1 = not at all, 7 = very much). Then participants were asked to rate how much of this review they believed was generated by AI (0% - 100%). The participants rated the six reviews separately. Lastly, participants answered an attention check question and demographic questions.

Results

First, we use our survey data to determine the relationship between perceptions of AI usage from our participants and detected AI usage from ZeroGPT. Results are presented in Table 10. For our pre-registered outcome of majority-AI-written, we find a highly significant, but modest, effect of a 6.8% higher probability of participant AI perceptions of majority AI usage when shown writing that was detected to use AI. Correspondingly, we find that the perceived percentage of AI writing is 6.5 percentage points higher for reviews that were detected by ZeroGPT to use AI, significant at the 0.1% level. The small magnitudes of the effects are in line with effect sizes found in earlier work showing high levels of imprecision in human perceptions of AI usage (Ma and Luo 2023).

We find further support for this hypothesis in regressions of detected AI usage and participant perceptions of review quality, presented in Table 11. We find a significant penalty in terms of perceived authenticity for reviews detected to be written with

Table 11: Effect of Detected AI Use on
Review Quality Perceptions

	Authentic	Helpful	Useful	Persuasive	Sincere
$\beta_{DetectedAI}$	-0.184*	-0.126	−0.147†	−0.150†	−0.158*
	(0.077)	(0.080)	(0.083)	(0.087)	(0.077)
$\beta_{DetectedAI\ X\ Perceived50\%+AI}$	-0.893***	-0.298*	-0.310**	-0.364**	-0.744***
	(0.109)	(0.115)	(0.119)	(0.125)	(0.109)
$\beta_{DetectedAI\ X\ NotPerceived50\%+AI}$	0.148†	-0.046	-0.071	-0.051	0.116
	(0.084)	(0.089)	(0.092)	(0.097)	(0.084)
N	1716	1716	1716	1716	1716

Notes: Effect of detected AI usage on review quality perceptions. Sample drawn from Prolific, stimuli drawn from field data (Yelp.com) with GPT use detected using ZeroGPT, stratified to include 25 reviews with detected AI usage and 25 reviews with no detected AI usage. Perceptions of AI usage solicited from survey participants after reviews were rated. † = 10% significance, * = 5% significance, ** = 1% significance, *** = 0.1% significance.

AI and insignificant negative point estimates across our other quality dimensions. However, when we inspect reviews that are detected to be written with AI and are perceived to be written with AI, following our pre-registered interaction analysis, we estimate a significant penalty across all quality dimensions.

Taken together with the results from experiments 1 and 2, this result suggests that the effect observed in the field may be driven by a penalty associated with perceptions of AI usage, and in particular a suspicion that AI-generated reviews are fake. Moreover, the results are suggestive that patterns in the field data may be affected by measurement error because that consumers' AI usage perceptions are also highly imprecise, with a small magnitude in participants' estimated ability to discriminate between detected AI-written reviews and detected human-written reviews. This suggests that our field result evidence may be a conservative lower bound of true effects due to attenuation bias. Indeed, if one combines our field evidence findings from Experiments 1 and 2, suggesting that the effect is driven by perceptions of AI usage in particular, with our estimates from Table 10, this suggests that our field evidence estimates may be attenuated by an entire order of magnitude (namely, $\frac{1}{0.068}$); ap-

plying an inverse scaling to our Yelp differences-in-differences results implies that a noiseless perception of AI usage would lead to a full standard deviation decline in perceived usefulness. Similarly, our evidence in Table 11 suggests that the penalty to perceived "authenticity" and "sincerity" is over twice as large as perceived penalties to "helpfulness", "usefulness" and "persuasiveness", implying that for contexts where authenticity is especially salient—for example, for online message boards, where authenticity of interactions with other humans is central to the utility that users derive from platform usage—the threat to utility may be even more profound (Kumar et al. 2017).

That said, considering the experimental evidence and the field evidence together, our results also provide support for a simple intervention that managers may implement that could plausibly reduce this AI usage penalty to zero: badges, modeled after "Vine Voice" at Amazon and "Elite" at Yelp, may provide a clear antidote to any deleterious effects from perceptions of AI usage on user-generated content websites, likely due to the fact that such badges verify reviews as non-fake. Managers may independently seek to consider such verification programs for their specific business and policymakers interested in preserving the public goods of the internet for society at larger may be advised to subsidize such "internet infrastructure" in order to combat the negative implications of widespread disaffection with online content.

5 Conclusion

There is little doubt that generative AI will transform the way we live in years to come. But even today, the technology is already spreading unhindered through our online ecosystem, with few checks or incentive structures to help guide its adoption. In the absence of such responsible structures for ensuring better use, many fear that generative AI will degrade the quality of content that we enjoy across a wide variety of domains, as workers use the technology as a time-saving device at the expense of output quality (Kobak et al. 2024). For online reviews in particular, many are concerned that both private businesses and public welfare may be incidentally harmed by the deterioration of trust in information found online, which may be felt especially acutely in the domain of online user ratings and reviews.

This study is the first, to the best of our knowledge, that provides a rigorous evidence that these fears are well-founded in the context of online reviews. Based on a sample of reviews of San Francisco restaurants on Yelp.com and Amazon.com product reviews for the "Camera, Photo and Video" category, we find that detected generative AI use significantly reduces the average perceived quality of reviewer output. Even when inspecting the same identified reviewer before the release of ChatGPT and after the release of ChatGPT, we find that a one standard deviation increase in detected AI usage leads to a 6.5% decrease in perceived Yelp.com review quality, as measured by the number of "useful" votes that the review receives.

We also find evidence that these effects differ across badged and non-badged reviews, with the effect entirely driven by non-badged reviews in our Yelp.com sample using our differences-in-differences specification. (Our sample is not large enough to make precise inferences for Amazon.com in our differences-in-differences specification, but point estimates are also approximately twice as large for non-badged versus badged reviews there.) This suggests that the effect is strongest in the absence of an expert badge.

We strengthen our evidence and explore the potential mechanisms more deeply in three pre-registered experiments and find clear evidence in favor of a negative penalty from perceptions of AI usage specifically, with null quality penalties found for actual AI writing: when we present online survey participants with reviews that are alternately described as AI-written and actually AI-written; described as human-written and actually AI-written; described as AI-written and actually human-written; and described as human-written and actually human-written, we find a strong negative effect on perceived quality of described AI use, but no effect of actual AI use. A follow-up study shows that this effect crucially hinges on the *non-verified* status of reviews, with effects largely vanishing when instead presenting participants with reviews that are either described as AI- or human-written but that are additionally described to be based on "verified customer experiences," suggesting that this penalty against perceived AI is driven in particular by perceptions of the likelihood that the review is fake. Finally, when we ask participants to evaluate 25 randomly selected reviews from our field data that were detected to involve AI usage and 25 randomly selected reviews that were not detected to involve AI usage, we find both 1) that

participant perceptions of AI usage is significantly predicted by detected AI usage, and 2) that reviews involving AI usage are rated as significantly lower-quality, in particular when they are detected to involve AI usage and are perceived as such by the given participant.

While this evidence lends strong support to the hypothesis that perceptible generative AI usage will harm the perceived quality of online reviews, the pattern of our results also suggests that businesses may not be helpless to fight this tide: for reviews that carry an expert badge like Yelp.com's "Elite" program or Amazon.com's "Vine Voice" program, generative AI use does not detectably harm perceived content quality, in line with our lab results on credibly "verified" reviews. Managers hoping to ameliorate these negative effects may seek to introduce or expand such programs on their platforms, and policymakers aiming to preserve social surplus may seek to subsidize credible validation systems for online reviews and ratings in order to maintain the value of crowdsourced information on consumer products (Reimers and Waldfogel 2021).

Generative AI will transform the world. At the same time, the internet is a public good, and unchecked generative AI usage poses a threat to the average perceived value of information found online. Maintaining credibility of online ratings and reviews is crucial both for managers to maintain profit and and for policymakers to maintain social welfare. While other research has found countless promising aspects of the generative AI revolution, discussions of generative AI's applications will be best served with clear evidence as to its potential drawbacks alongside its potential benefits, to help guide this transformation towards a more productive, more profitable, and higher-welfare world.

References

Bellaiche, L., R. Shahi, M.H. Turpin et al., "Humans versus AI: whether and why we prefer human-created compared to AI-created artwork," *Cognitive Research: Principles and Implications*, 2023, *8* (42). 7, 25

Boyd, Ryan L., Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker, "The Development and Psychometric Properties of LIWC-22," Technical Report, University of Texas at Austin 2022. LIWC Manual. 17, 51, 52

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield et al., "Toward trustworthy AI development: Mechanisms for supporting verifiable claims," *arXiv preprint arXiv:2004.07213*, 2020. 6

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond, "Generative AI at Work," April 2023, (31161). 1, 6, 25

Burtch, Gordon, Dokyun Lee, and Zhichen Chen, "The Consequences of Generative AI for UGC and Online Community Engagement," May 1 2023. Available at SSRN: https://ssrn.com/abstract=4521754 or http://dx.doi.org/10.2139/ssrn.4521754. 6

Callaway, Brantly and Pedro H.C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230. Themed Issue: Treatment Effect 1. 16

Capraro, Valerio, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-François Bonnefon, Pablo Brañas-Garza, Luigi Butera, Karen M Douglas, Jim A C Everett, Gerd Gigerenzer, Christine Greenhow, Daniel A Hashimoto, Julianne Holt-Lunstad, Jolanda Jetten, Simon Johnson, Werner H Kunz, Chiara Longoni, Pete Lunn, Simone Natale, Stefanie Paluch, Iyad Rahwan, Neil Selwyn, Vivek Singh, Siddharth Suri, Jennifer Sutcliffe, Joe Tomlinson, Sander van der Linden, Paul A M Van Lange, Friederike Wall, Jay J Van Bavel, and Riccardo Viale, "The impact of generative artificial intelligence on socioeconomic inequalities and policy making," *PNAS Nexus*, 06 2024, *3* (6), pgae191. 6

Castelo, Noah, Maarten W. Bos, and Donald R. Lehmann, "Task-dependent algorithm aversion," *Journal of Marketing Research*, 2019, *56* (5), 809–825. 1, 7

Chen, Jiafeng and Jonathan Roth, "Logs with Zeros? Some Problems and Solutions," *The Quarterly Journal of Economics*, May 2024, *139* (2), 891–936. Published online:

14 December 2023. 17

Chevalier, Judith A. and Dina Mayzlin, "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 2006, *43* (3), 345–354. 1, 7

Chiarella, Salvatore G., Giulia Torromino, Dionigi M. Gagliardi, Dario Rossi, Fabio Babiloni, and Giulia Cartocci, "Investigating the negative bias towards artificial intelligence: Effects of prior assignment of AI-authorship on the aesthetic appreciation of abstract paintings," *Computers in Human Behavior*, 2022, *137*, 107406. 7

Chung, Jaeyeon, "AI Luddites: Consumers Penalize Creative Work Output Generated by Artificial Intelligence," 10 2023. 25

Clare, Carl, Gillian Wright, Peter Sandiford, and Alberto Paucar-Caceres, "Why should I believe this? Deciphering the qualities of a credible online customer review," *Journal of Marketing Communications*, 02 2016, *24*, 1–20. 7

Craciun, Georgiana and Keryn Moore, "Credibility of negative online product reviews: Reviewer gender, reputation and emotion effects," *Computers in Human Behavior*, 2019, *97*, 104–115. 7

Dargnies, Marie-Pierre, Rustamdjan Hakimov, and Dorothea Kübler, "Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence," *Management Science*, 2024. 7

Díaz-Rodríguez, Natalia, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera, "Connecting the dots in trustworthy artificial intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Information Fusion*, 2023, *99*, 101896. 6

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *Journal of Experimental Psychology: General*, 2015, *144* (1), 114–126. 7

Dong, Beibei, M. Li, and K. Sivakumar, "Online review characteristics and trust: A cross-country examination," *Decision Sciences*, 2019, *50* (3), 537–566. 8

Doshi, Anil Rajnikant and Oliver Hauser, "Generative Artificial Intelligence Enhances Creativity but Reduces the Diversity of Novel Content," August 8 2023. Available at SSRN: https://ssrn.com/abstract=4535536 or http://dx.doi.org/10.2139/ssrn.4535536. 1

Freitas, Julian De, Saileena Agarwal, Bernd Schmitt, and Nick Haslam, "Psychological factors underlying attitudes toward AI tools," *Nature Human Behaviour*, 2023, *7* (11), 1845–1854. 7

He, Sherry, Brett Hollenbeck, Gijs Overgoor, and Ali Tosyali, "Detecting fake-review buyers using network structure: Direct evidence from Amazon," *Proceedings of the National Academy of Sciences*, 2022, *119* (47), e2211932119. Edited by Avi Goldfarb, University of Toronto, Canada; received July 13, 2022; accepted October 10, 2022 by Editorial Board Member Mark Granovetter. 8

Hui, Xiang, Oren Reshef, and Luofeng Zhou, "The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market," July 31 2023. Available at SSRN: https://ssrn.com/abstract=4527336 or http://dx.doi.org/10.2139/ssrn.4527336. 6

Hutto, Clayton and Eric Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in "Proceedings of the International AAAI Conference on Web and Social Media," Vol. 8 May 2014, pp. 216–225. 11

Jr, C.B. Horton, M.W. White, and S.S. Iyengar, "Bias against AI art can enhance perceptions of human creativity," *Scientific Reports*, 2023, *13*, 19001. 7

Kaur, Davinder, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi, "Trustworthy artificial intelligence: a review," *ACM computing surveys (CSUR)*, 2022, *55* (2), 1–38. 6

Kobak, Dmitry, Rita Márquez, Emőke Ágnes Horvát, and Jan Lause, "Delving into ChatGPT usage in academic writing through excess vocabulary," 06 2024. 35

Kowald, Dominik, Sebastian Scher, Viktoria Pammer-Schindler, Peter Mullner, Kerstin Waxnegger, Lea Demelius, Angela Fessl, Markus Toller, Inti Gabriel Mendoza Estrada, Ilija Simic et al., "Establishing and evaluating trustworthy AI: Overview and research challenges," *Frontiers in Big Data*, 2024, *7*, 1467222. 6

Kumar, Srijan, Justin Cheng, Jure Leskovec, and V. S. Subrahmanian, "An Army of Me: Sockpuppets in Online Discussion Communities," in "Proceedings of the 26th International Conference on World Wide Web (WWW '17)" International World Wide Web Conferences Steering Committee Perth, Australia 2017, pp. 857–866. arXiv:1703.07355 [cs.SI]. 35

Li, Bo, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou, "Trustworthy AI: From principles to practices," *ACM Computing Surveys*, 2023,

55 (9), 1–46. Published online 2021. 6

Li, Yuheng, Lele Sha, Lixiang Yan, Jionghao Lin, Mladen Raković, Kirsten Galbraith, Kayley Lyons, Dragan Gašević, and Guanliang Chen, "Can large language models write reflectively," *Computers and Education: Artificial Intelligence*, 2023, *4*, 100140. 1

Longoni, Chiara, Andrea Bonezzi, and Carey K. Morewedge, "Resistance to medical artificial intelligence," *Journal of Consumer Research*, 2019, *46* (4), 629–650. 1, 7

Ma, Liye and Lan Luo, "Beyond Fake or Genuine – The Effect of Large Language Models (LLMs) on the Content and Sentiment of Product Reviews," *USC Marshall School of Business Research Paper Sponsored by iORB*, July 14 2023, (Forthcoming). Available at SSRN: https://ssrn.com/abstract=4511025. 4, 10, 33

Mudambi, Susan M and David Schuff, "Research note: What makes a helpful online review? A study of customer reviews on Amazon. com," *MIS quarterly*, 2010, pp. 185–200. 1

Noy, Shakked and Whitney Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, 2023, *381* (6654), 187–192. 1

Otis, Nicholas G., Rowan Clarke, Solène Delecourt, David Holtz, and Rembrand Koning, "The Uneven Impact of Generative AI on Entrepreneurial Performance," Working Paper 24-042, Harvard Business School December 2023. 6

Reimers, Imke and Joel Waldfogel, "Digitization and Pre-purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings," *American Economic Review*, June 2021, *111* (6), 1944–71. 1, 37

Roose, Kevin, "GPT-4 Is Exciting and Scary," *The New York Times*, March 2023. 1

Rosario, Alexandra Babić, Francesca Sotgiu, Kristine De Valck, and Tammo H.A. Bijmolt, "The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors," *Journal of Marketing Research*, 2016, *53* (3), 297–318. 1

Rosario, Ana Babić, Kristine de Valck, and Francesca Sotgiu, "Conceptualizing the electronic word-of-mouth process: What we know and need to know about eWOM creation, exposure, and evaluation," *Journal of the Academy of Marketing Science*, 2020, *48*, 422–448. 7

shin Lim, Young and Brandon Van Der Heide, "Evaluating the Wisdom of Strangers: The Perceived Credibility of Online Consumer Reviews on Yelp," *Journal of Computer-*

Mediated Communication, January 2015, *20* (1), 67–82. 7

Thiebes, Scott, Sebastian Lins, and Ali Sunyaev, "Trustworthy artificial intelligence," *Electronic Markets*, 2021, *31* (2), 447–464. 6

Wu, Yan, E. W. Ngai, Pui Wu, and Chihang Wu, "Fake online reviews: Literature review, synthesis, and directions for future research," *Decision Support Systems*, 2020, *132*, 113280. 8

Yeomans, Mike, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg, "Making sense of recommendations," *Journal of Behavioral Decision Making*, 2019, *32* (4), 403–414. 7

A1 Quantity Effects

We here present analysis of the effect of detected AI usage on quantity of reviews written. Results are presented in Appendix Tables A1 and A2. In both settings, we see strikingly similar patterns, both in the pre- and post-ChatGPT release period: prior to ChatGPT's release, reviewers who were detected as more frequently using ChatGPT in review production were associated with significantly lower quantity of production, but after ChatGPT's November 2022 release, that association flips and becomes significantly positive. The negative pre-period estimates suggest that producing reviews that looked "ChatGPT-like", including the use of polished phrases and complete sentences, may have required more labor prior to the release of ChatGPT. The coefficient estimates suggest that a one standard deviation increase in GPT adoption leads to a 1.2% increase in reviewer output quantity in the post-ChatGPT release period for Yelp.com, or a 1.5% increase in reviewer output quantity for Amazon.com.

Heterogeneity by Reviewer Badge Presence

To investigate how this effect may be moderated by badge presence, we stratify the sample between badged and non-badged reviewers based on each respective website's 'expert reviewer badge' and estimate our baseline specification for both badged and non-badged categories.[21] Results are presented in Appendix Tables A3 and A4. As with our quality estimates, we find that our observational relationships remain significant and do not change direction across all subcategories in this observational setting, suggesting that the effect is fairly broad-based across groups. For Amazon.com reviews, point estimates are quite similar across badged and non-badged reviewers, while for Yelp.com reviews, quantity effects appear to be much more pronounced for badged reviewers. That said, this specification does not control for potential differential selection into AI usage that also differs across groups; with this in mind, we turn to our differences-in-differences specification again.

Table A1: Effect of Detected AI Use on Log Number of Yelp Reviews

$\beta_{GPT-pre}$	-0.011***	-0.013***
	(0.003)	(0.003)
$\beta_{GPT-post}$	0.016***	0.012***
	(0.003)	(0.003)
Controls		X
N	44432	44432

Table A2: Effect of Detected AI Use on Log Number of Amazon Reviews

$\beta_{GPT-pre}$	-0.007***	-0.009***
	(0.002)	(0.002)
$\beta_{GPT-post}$	0.021***	0.014***
	(0.002)	(0.002)
Controls		X
N	52176	52176

Notes: Effect of adoption of ChatGPT in review production on log review quantity produced per reviewer, per period. Sample based on Yelp reviews for restaurants in San Francisco area and reviews for products in "Camera, Photo and Video" category of Amazon, with GPT use detected using ZeroGPT. Controls include average star rating of review, wordcount of review, valence of review, and price level dummies, averaged at the reviewer-by-period level. Outcome variables standardized to express effects in terms of standard deviation shifts. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

Table A3: Effect of Detected AI Use on
Log Number of Yelp Reviews:
"Elite" vs. non-"Elite"

	"Elite"		Non-"Elite"	
$\beta_{GPT-pre}$	-0.027**	-0.026**	-0.009***	-0.009***
	(0.009)	(0.009)	(0.002)	(0.002)
$\beta_{GPT-post}$	0.037***	0.037***	0.011***	0.010***
	(0.010)	(0.010)	(0.002)	(0.002)
Controls		X		X
N	9644	9644	34788	34788

Table A4: Effect of Detected AI Use on
Log Number of Amazon Reviews

	"Vine Voice"		Non-"Vine Voice"	
$\beta_{GPT-pre}$	-0.018*	-0.023**	-0.006***	-0.007***
	(0.007)	(0.007)	(0.002)	(0.002)
$\beta_{GPT-post}$	0.014*	0.011	0.017***	0.011***
	(0.007)	(0.007)	(0.002)	(0.002)
Controls		X		X
N	6208	6208	45968	45968

Notes: Effect of adoption of ChatGPT in review production on log review quantity produced per reviewer, per period. Sample based on Yelp reviews for restaurants in San Francisco area and reviews for products in "Camera, Photo and Video" category of Amazon, with GPT use detected using ZeroGPT. Controls include average star rating of review, wordcount of review, valence of review, price level dummies, averaged at the reviewer-by-period level. Outcome variables standardized to express effects in terms of standard deviation shifts. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

Differences-in-Differences Estimates

To account for possible selection effects and to control against other individual-specific or period-specific confounds, we collapse our data to the reviewer-by-period level and estimate a differences-in-differences model of the effect of AI adoption on the log number of reviews produced per reviewer in each period, following equation 3. Results are presented in Appendix Tables A5 and A6. Overall, we find no significant effects of detected AI usage on the quantity of reviews produced in either setting under this specification. It appears that the observed effect in the baseline specification is driven primarily by across-author variation in our sample, and we are not able to rule out differential selection as a possible explanation for our observed quantity effects.

[21] For more detail on these categories, see section 4.1.1.

Table A5: Effect of Detected AI Use on Log Number of Yelp Reviews: Differences-in-Differences

	(1)	(2)
$\beta_{GPT-pre}$	0.001	0.000
	(0.010)	(0.010)
$\beta_{GPT-post}$	0.005	0.004
	(0.010)	(0.010)
δ_j	X	X
γ_t	X	X
Controls		X
N	9682	9682

Table A6: Effect of Detected AI Use on Log Number of Amazon Reviews: Differences-in-Differences

	(1)	(2)
$\beta_{GPT-pre}$	-0.002	-0.002
	(0.012)	(0.012)
$\beta_{GPT-post}$	-0.002	-0.001
	(0.021)	(0.021)
δ_j	X	X
γ_t	X	X
Controls		X
N	2578	2578

Notes: Difference-in-differences estimates of the effect of adoption of ChatGPT in review production on review quantity produced per reviewer, per 11 months. Sample based on Yelp reviews for restaurants in the San Francisco area and reviews for products in the "Camera, Photo and Video" category of Amazon, with GPT use detected using ZeroGPT. Controls include average star rating of review, word count of review, valence of review, and price level dummies, averaged at the reviewer-by-period level. Outcome variables are standardized to express effects in terms of standard deviation shifts. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

A2 Validity Check: Pre-Period Comparison of Authors Detected to Use AI in Post-Period

In this section, we present our regression results on pre-period differences between authors detected to use AI in the post-period and authors not detected to use AI in the post-period. Results are presented in A7. In this table, we define "GPT-post-author" as an indicator for a reviewer who is detected to use GPT in the post-period. $\beta_{GPT-post-author}$ measures the difference in average usefulness (or helpfulness) of reviews for these reviewers in the preperiod, compared to those who are not detected to use GPT in the post-period. The regression is author-level and restricted to the pre-period sample. We find null statistical differences for both our Yelp and Amazon samples, suggesting that pre-period differences between authors who later adopt GPT is highly unlikely to explain the post-period decline in perceived quality for their reviews.

Table A7: Pre-period Differences in Average Quality
for Authors with Post-Period GPT Use

	Yelp ("Usefulness")		Amazon ("Helpfulness")	
$\beta_{GPT-post-author}$	0.154	0.118	-0.002	-0.018
	(0.099)	(0.096)	(0.068)	(0.063)
Controls		X		X
N	4841	4841	1289	1289

Notes: Each column reports results from regressions of standardized pre-period review quality on an indicator for authors who later produce GPT-detected text. Controls include (author-level) average review star rating, word count, valence, and price category densities. The coefficient $\beta_{GPT-post-author}$ captures average pre-period differences between authors who use GPT in the post period versus those who do not. Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

A3 Only Author Fixed Effects

We also include an additional robustness checks to evaluate an alternative model for controlling for author effects. Here, we present results using not a two-period differences-in-differences design, but with a simple addition of an author fixed effect term. This is an alternative specification to identify effects from within-author variation, but to allow us to include observations from authors who may be only observed in a single period. This specification is in many ways less stringent than our preferred two-period differences-in-differences specification, which absorbs period fixed effects and weights each author equally, but nonetheless results are qualitatively identical to our differences-in-differences estimates. Estimates are presented in Appendix Tables A8 and A9. For our Yelp sample, we continue to see very strong and significant effects, but for our Amazon sample, as we observe very few multi-review authors (around 90,000 authors produced the observed 125,000 reviews for Amazon, compared to around 10,000 authors for Yelp) the effects are not significantly identified from zero when using only within-author variation, similar to our main differences-in-differences results.

Table A8: Effect of Detected AI Use on
Yelp.com Review "Useful" Votes: Author FE

$\beta_{GPT-pre}$	-0.024	-0.021
	(0.021)	(0.020)
$\beta_{GPT-post}$	-0.087***	-0.091***
	(0.018)	(0.018)
δ_j	X	X
Controls		X
N	79233	79233

Table A9: Effect of Detected AI Use on
Amazon.com Review "Helpful" Votes: Author FE

$\beta_{GPT-pre}$	-0.019	0.025
	(0.040)	(0.038)
$\beta_{GPT-post}$	-0.027	-0.031
	(0.031)	(0.030)
δ_j	X	X
Controls		X
N	124513	124513

Notes: Effect of adoption of ChatGPT in review production on review quality. Sample based on Yelp reviews for restaurants in San Francisco area and reviews for products in "Camera, Photo and Video" category of Amazon, with GPT use detected using ZeroGPT. Controls include star rating of review, wordcount of review, average valence of review, price level dummies, and for Yelp restaurants, restaurant type dummies. Outcome variables standardized to express effects in terms of standard deviation shifts in quality. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

A4 Explaining Amazon Pre-period Differences: Writing Style Controls

As described in the main text, we examine the extent to which writing style may account for the observed positive impact of "detected GPT" writing in reviews in the pre-period, per our hypothesis that well-written reviews may have been more likely to be detected as involving GPT usage in the pre-period, accounting for that earlier positive association. To examine this, we include writing style measures based on the major categories of writing style measured using the LIWC dictionary of Boyd et al. (2022) (interacted with pre- and post-period dummies to allow for the effect of writing style to shift after GPT is released) as controls, and examine how effects differ.

Table A10: Effect of Detected AI Use on
Amazon.com Review "Helpful" Votes:
With Linguistic Style Controls

$\beta_{GPT-pre}$	0.165***	0.080***	0.040*
	(0.019)	(0.018)	(0.018)
$\beta_{GPT-post}$	-0.149***	-0.142***	-0.076***
	(0.018)	(0.017)	(0.017)
Baseline Controls		X	X
Writing Style Controls			X
N	124513	124513	124513

Notes: Effect of adoption of ChatGPT in review production on review quality. Sample based on Yelp reviews for restaurants in San Francisco area and reviews for products in "Camera, Photo and Video" category of Amazon, with GPT use detected using ZeroGPT. Controls include star rating of review, wordcount of review, average valence of review, price level dummies. Linguistic style controls include the full set of stylistic category density measures from Boyd et al. (2022) including "Tone", "WPS", "Bigwords", "DIC", "Drives", "Cognition", "Affect", "Social", "Culture", "Lifestyle", "Physical", "Perception", "Conversation", interacted with pre- and post-period dummies to allow for differential effects of writing style in either period. See Boyd et al. (2022) for further details on these style categories. Outcome variables standardized to express effects in terms of standard deviation shifts in quality. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

.

Results are presented in Table A10. In line with our hypothesis, we find that including writing style controls does meaningfully attenuate the pre-period positive association. The post-period effect remains highly significant and negative but is also attenuated. We do not include these controls in our main regression out of concerns that writing style is a "bad control" in our setting, much more so that the other controls that we include in our baseline model, since we posit that writing style is the primary mechanism by which readers determine perceptions of AI usage, which we identify as the main channel of effects, based on the evidence from our experiments. Nonetheless, we suggest that this pattern of results is reassuring that the positive association between detected-GPT reviews and quality in the pre-period is driven

by a "well-written" writing style that later becomes associated with GPT writing, reinforcing our interpretation that the post-period effects may be best interpreted as conservative lower-bounds.

A5 Details on Experiments

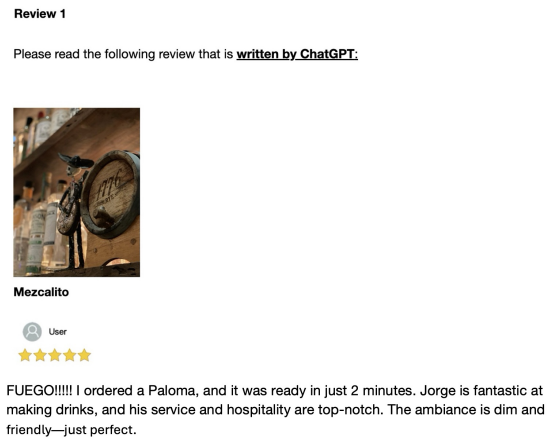
Example Stimulus and Instruction Details for Experiments 1 and 2

We here present screenshot examples of the specific stimuli displayed to online survey participants in each condition for Experiment 1. Each participant was shown two such stimuli out of a list of 25 reviews, and asked to rate each of the two reviews on a Likert scale of 1-7 across five pre-registered quality dimensions: authenticity, helpfulness, usefulness, persuasiveness, and sincerity. Participants were also instructed on the following before they read the reviews and based on the communicated writing agent (AI vs. human) conditions they were assigned to: "Please read the following review that is written by ChatGPT" or "Please read the following review that is written by the customer herself."

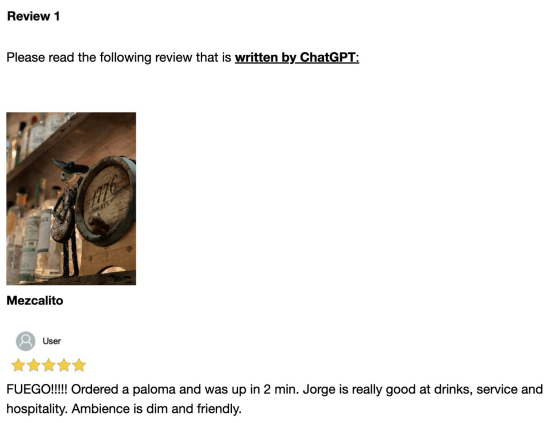
Full details of every text stimulus used in this experiment are available in our online supplemental files.

Figure A1: Experiment 1: Example Stimuli

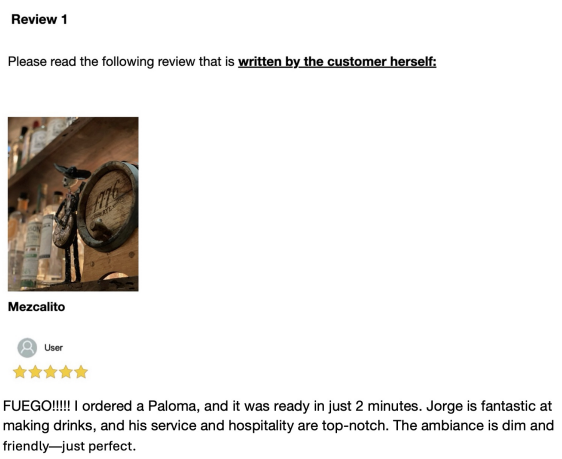
(a) "AI-written" x "informed-AI" condition



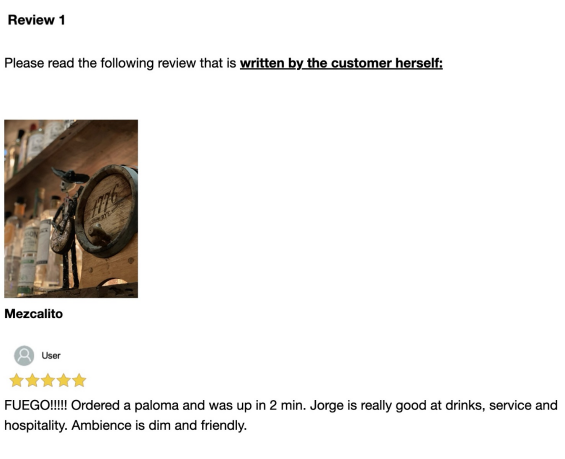
(b) "Human-written" x "informed-AI" condition



(c) "AI-written" x "informed-human" condition



(d) "Human-written" x "informed-human" condition



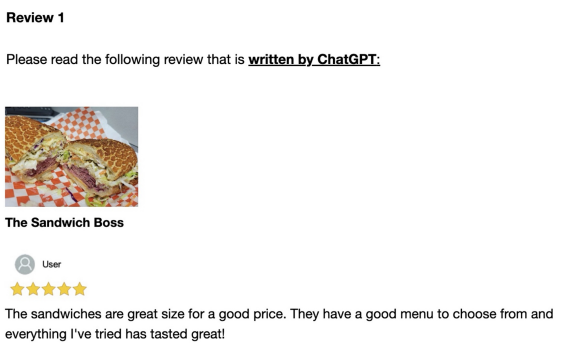
Notes: Example stimuli for experiment 1, informed AI (a and b) and informed human (c and d) conditions. Stimuli text is written by AI in AI-written conditions (a and c), and written by human in human-written conditions (b and d). Survey drawn from field data sample of Yelp.com reviews.

For experiment 2, we used the same text stimulus as in experiment 1, but amended the instructions before presenting the stimuli. Participants were also instructed on the following before they read the reviews and based on the communicated writing agent (AI vs. human) conditions and the verification conditions they were assigned to: (1) communicated agent: "Please read the following review that is written by ChatGPT" or "Please read the following review that is written by the customer

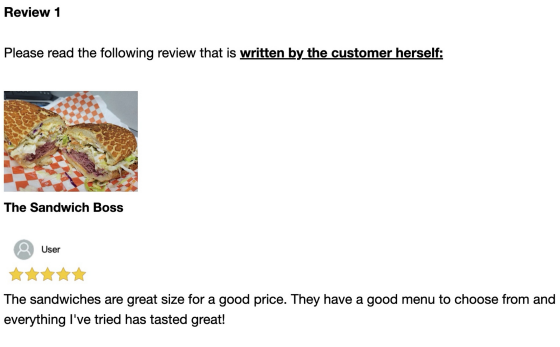
herself" or "Please read the following review that is written by the customer herself and edited by ChatGPT"; followed by (2) verification: "based on a verified customer experience" or no information on verification. We here present screenshot examples of the specific stimuli displayed in each condition for Experiment 2. Full details of every text stimulus used in this experiment are available in our online supplemental files.

Figure A2: Experiment 2: Example Stimuli for All Conditions

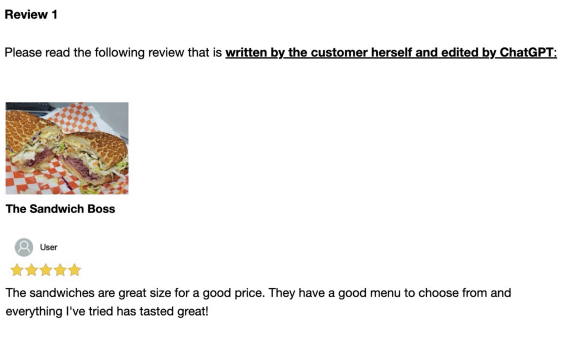
(a) "AI-written" condition



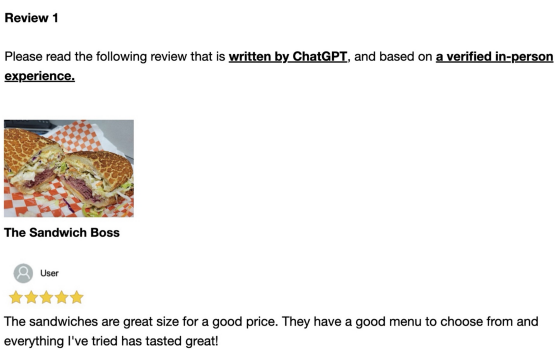
(b) "Human-written" condition



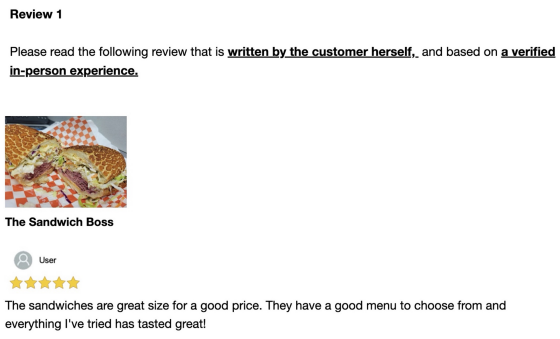
(c) "AI-edit" condition



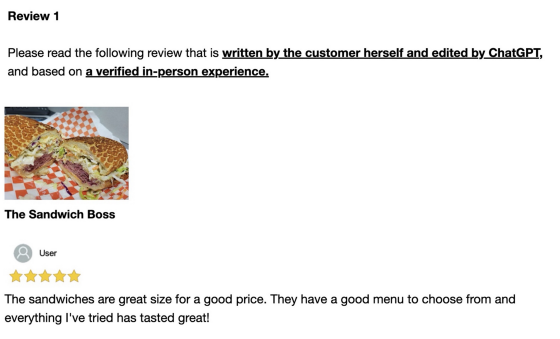
(d) Verified "AI-written"



(e) Verified "human-written"



(f) Verified "AI-edit"



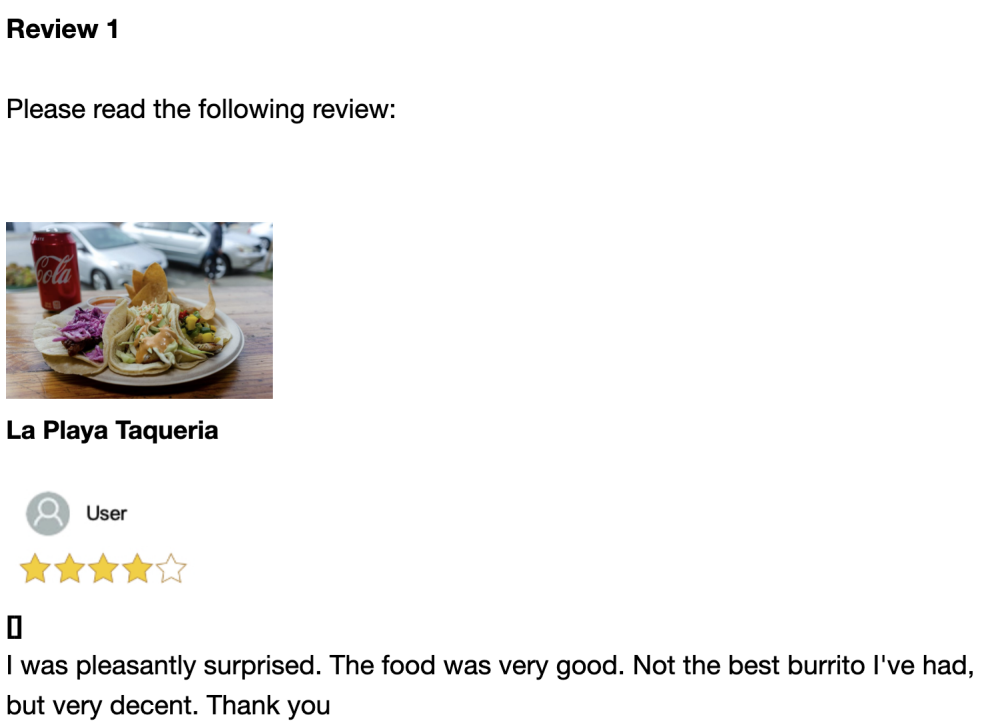
Notes: Example stimuli for all Experiment 2 conditions. Survey drawn from field data sample of Yelp.com reviews.

Example Stimulus for Experiment 3

We here present a screenshot example of the specific stimuli displayed to online survey participants for Experiment 3. Each participant was shown six such stimuli and

asked to rate it on a Likert scale of 1-7 across five pre-registered quality dimensions: authenticity, helpfulness, usefulness, persuasiveness, and sincerity, as well as the % of the review that they perceived to involve AI usage. The regression analysis, including the regression with the interaction with majority-perceived-AI-usage, were all pre-registered. Full details of every text stimulus used in this experiment are available in our online supplemental files.

Figure A3: Experiment 3: Example Stimulus



Notes: Example stimulus for experiment 3. Survey drawn from field data sample of Yelp.com reviews.

Additional Analysis in Experiment 2

As pre-registered, we reported results that excluded participants in the "verified" conditions who did not believe our verification. Next, we report results that include these participants in the analysis.

Results are presented in Table A11. Compared to the reviews that were communicated as written by humans, conditions that communicated AI usage, either written by AI or edited by AI, were rated as lower-quality across all the dimensions. We also found the effects being attenuated in the AI-edit condition: compared to reviews that were not verified and involved AI usage, reviews that were verified showed smaller negative effects across the dimensions. The negative effects in the "verified" conditions were more significant than the results reported in 9, which is expected as those who were not influenced by the verification interventions are likely to show a similar level of negative effects on AI-written reviews even in the "verified" conditions.

Table A11: Effect of Communicated AI Usage and Verified Experience on Review Quality Perceptions (No Exclusion)

	Authentic	Helpful	Useful	Persuasive	Sincere
$\beta_{CommunicatedAI\ X\ Verify}$	-0.778^{***}	-0.360^{*}	-0.252	-0.463^{**}	-0.606^{***}
	(0.161)	(0.167)	(0.161)	(0.174)	(0.157)
$\beta_{CommunicatedAI\ X\ NoVerify}$	-1.348^{***}	-0.843^{***}	-0.758^{***}	-0.884^{***}	-1.045^{***}
	(0.160)	(0.166)	(0.160)	(0.173)	(0.156)
$\beta_{CommunicatedAIEdit\ X\ Verify}$	-0.603^{***}	-0.422^{*}	-0.417^{*}	-0.550^{**}	-0.492^{**}
	(0.161)	(0.167)	(0.161)	(0.174)	(0.157)
$\beta_{CommunicatedAIEdit\ X\ NoVerify}$	-0.708^{***}	-0.641^{***}	-0.620^{***}	-0.703^{***}	-0.696^{***}
	(0.160)	(0.166)	(0.160)	(0.172)	(0.156)
$\beta_{CommunicatedHuman\ X\ Verify}$	-0.089	-0.019	-0.004	$-0.287^{\dagger}$	0.042
	(0.162)	(0.168)	(0.162)	(0.175)	(0.158)
$\beta_{CommunicatedHuman\ X\ NoVerify}$	.	.	.	.	.
(Excluded Category)	(.)	(.)	(.)	(.)	(.)
N	1,174	1,174	1,174	1,174	1,174

Notes: Effect of different combinations of communicated AI usage and verification of experiences on review quality perceptions. Sample drawn from Prolific and stimuli drawn from field data (Yelp.com) reviews that were detected to involve no GPT use, with GPT use detected using ZeroGPT. Usage was communicated explicitly to be either "written by the customer herself", "written by the customer herself and edited by ChatGpt", or "written by ChatGPT". Verification of experience was communicated by explicitly to be "based on a verified in-person experience".† = 10% significance, * = 5% significance, ** = 1% significance, *** = 0.1% significance.

Deviation from Pre-Analysis Plan: Additional Analysis in Experiment 3

While our primary analyses of review quality were pre-registered, we added one analysis, based on pre-registered variables, that was not in our original pre-analysis plan. Specifically, for Experiment 3, while we noted that we would collect participant perceptions of majority-AI usage as part of a pre-registered interaction term specification, we did not pre-register the analysis of participant perceptions of majority AI-usage and percentage of AI usage against detected AI usage using ZeroGPT, presented in Table 10. After running the experiment, this regression was easily accomplished with our collected data and filled an important gap in our examination, and so we included this analysis in addition to the pre-registered regressions of quality perceptions both with and without interactions.

A6 Additional Tables and Figures

Table A12: Effect of Detected AI Use on
Yelp.com Review Log("Useful" Votes)

$\beta_{GPT-pre}$	-0.048	-0.027
	(0.042)	(0.040)
$\beta_{GPT-post}$	-0.184***	-0.212***
	(0.047)	(0.045)
Controls		X
N	33562	33562

Table A13: Effect of Detected AI Use on
Amazon.com Review Log("Helpful" Votes)

$\beta_{GPT-pre}$	0.339***	0.253***
	(0.027)	(0.024)
$\beta_{GPT-post}$	-0.440***	-0.375***
	(0.034)	(0.029)
Controls		X
N	72985	72985

Notes: Effect of adoption of ChatGPT in review production on review quality, using log(x) functional form. Sample based on Yelp reviews for restaurants in San Francisco area and reviews for products in "Camera, Photo and Video" category of Amazon, with GPT use detected using ZeroGPT. Controls include star rating of review, wordcount of review, average valence of review, price level dummies, and for Yelp restaurants, restaurant type dummies. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

Table A14: Effect of Detected AI Use on Review "Funny"-ness: Yelp.com

$\beta_{GPT-pre}$	-0.035	-0.036
	(0.038)	(0.037)
$\beta_{GPT-post}$	-0.080*	-0.134***
	(0.032)	(0.032)
Controls		X
N	79233	79233

Table A15: Effect of Detected AI Use on Review "Cool"-ness: Yelp.com

$\beta_{GPT-pre}$	-0.023	-0.029
	(0.038)	(0.037)
$\beta_{GPT-post}$	-0.085**	-0.148***
	(0.032)	(0.032)
Controls		X
N	79233	79233

Notes: Effect of adoption of ChatGPT review production on review quality for alternative quality measurements. Sample based on Yelp reviews for restaurants in San Francisco area and reviews for products in "Camera, Photo and Video" category of Amazon, with GPT use detected using ZeroGPT. Controls include star rating of review, wordcount of review, average valence of review, price level dummies, and for Yelp restaurants, restaurant type dummies. Outcome variables standardized to express effects in terms of standard deviation shifts in quality. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

Table A16: Effect of Detected AI Use on
Yelp.com Review "Useful" Votes:
"Elite" vs. non-"Elite"

	"Elite"		Non-"Elite"	
$\beta_{GPT-pre}$	-0.031	-0.020	0.039	0.038
	(0.061)	(0.059)	(0.047)	(0.047)
$\beta_{GPT-post}$	-0.171***	-0.203***	-0.060	-0.091*
	(0.048)	(0.047)	(0.044)	(0.043)
Controls		X		X
N	30733	30733	48500	48500

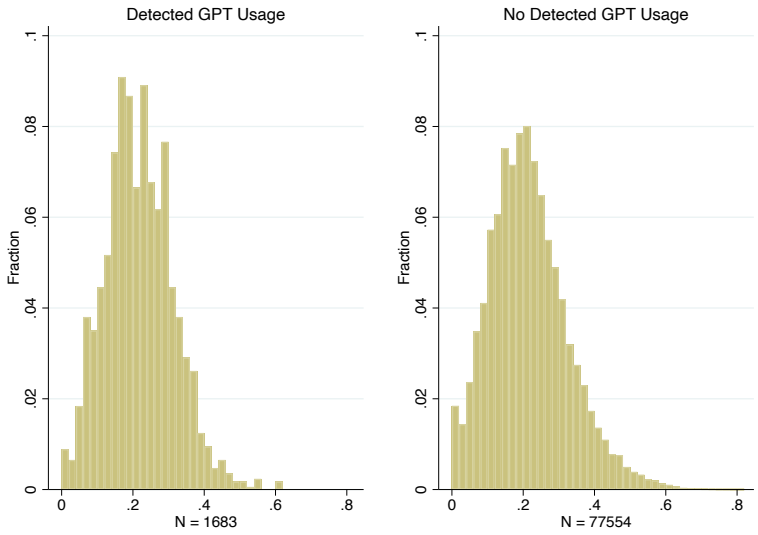
Table A17: Effect of Detected AI Use on
Amazon.com Review "Helpful" Votes:
"Vine Voice" vs. non-"Vine Voice"

	"Vine Voice"		Non-"Vine Voice"	
$\beta_{GPT-pre}$	0.168***	0.078	0.161***	0.077***
	(0.051)	(0.047)	(0.021)	(0.021)
$\beta_{GPT-post}$	-0.096**	-0.064*	-0.146***	-0.133***
	(0.029)	(0.027)	(0.022)	(0.020)
Controls		X		X
N	15529	15529	108984	108984

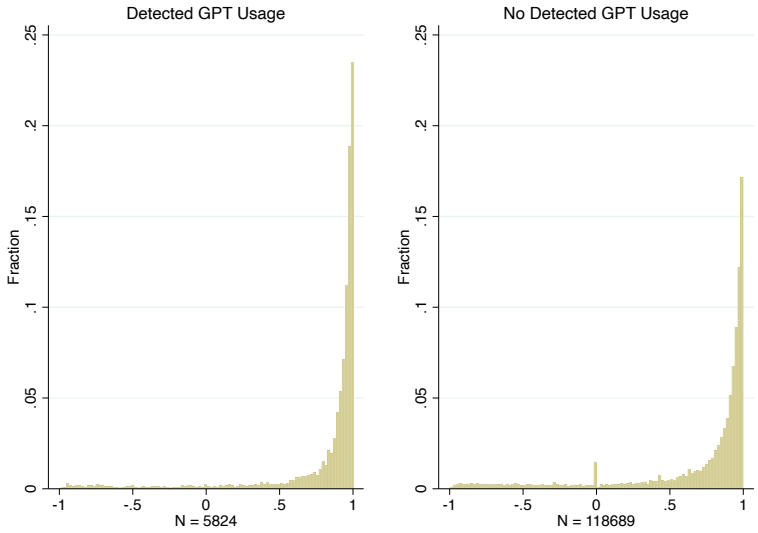
Notes: Effect of adoption of ChatGPT in review production on review quality, stratified by reviewer category. Sample based on Yelp reviews for restaurants in San Francisco area and reviews for products in "Camera, Photo and Video" category of Amazon, with GPT use detected using ZeroGPT. Controls include star rating of review, wordcount of review, average valence of review, price level dummies, and for Yelp restaurants, restaurant type dummies. Outcome variables and covariates standardized to express effects in terms of standard deviation shifts. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

Figure A4: Valence Distribution:
Human versus AI

(a) Yelp.com



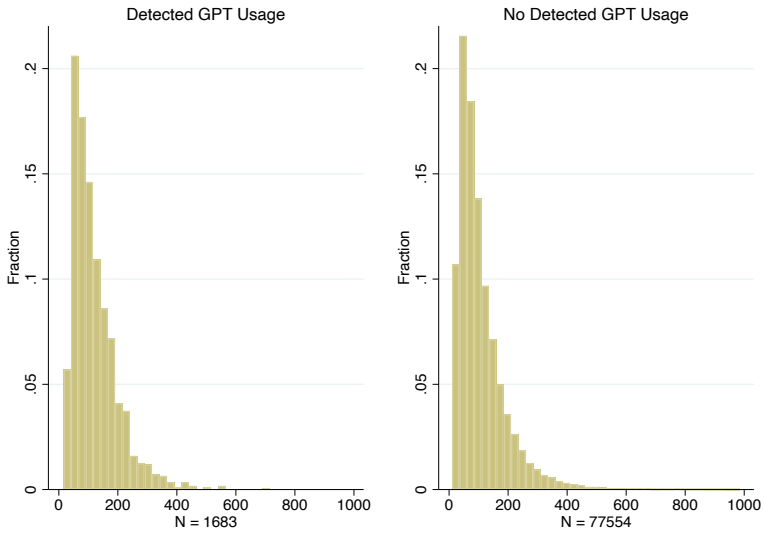
(b) Amazon.com



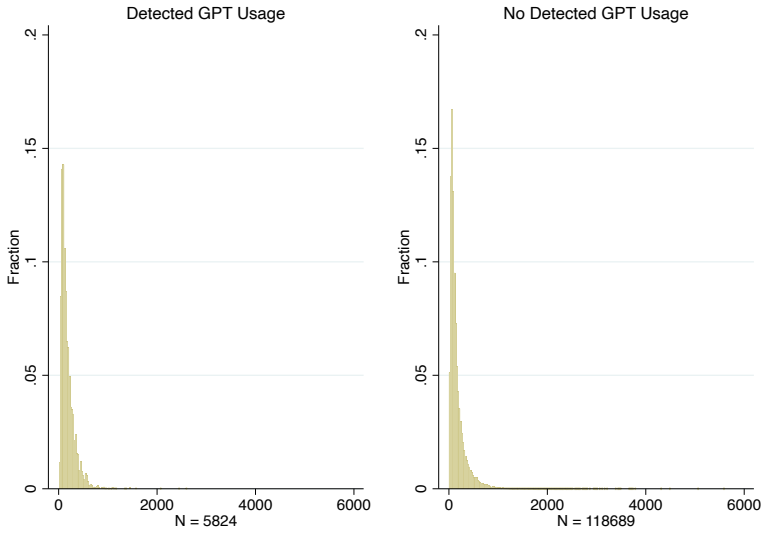
Notes: Valence distribution of reviews, stratified by detected AI usage. Y-axis of each panel shows the fraction of reviews at each valence, while X-axis shows the associated valence. Use of generative AI scored using ZeroGPT.com. Sample based on 79,233 reviews gathered from Yelp.com for the universe of restaurants in San Francisco, and 124,513 reviews gathered from Amazon.com for the "Camera, Photo and Video" category.

Figure A5: Wordcount Distribution:
Human versus AI

(a) Yelp.com



(b) Amazon.com



Notes: Wordcount distribution of reviews, stratified by detected AI usage. Y-axis of each panel shows the fraction of reviews at each wordcount, while X-axis shows the associated wordcount. Use of generative AI scored using ZeroGPT.com. Sample based on 79,233 reviews gathered from Yelp.com for the universe of restaurants in San Francisco, and 124,513 reviews gathered from Amazon.com for the "Camera, Photo and Video" category.