Working Paper

Generative AI and User-Generated Content: Evidence from Online Reviews

Working Paper Number 15

Samsun Knight

Yakov Bart

WORKING PAPER

Working Paper No. 15

Generative AI and User-Generated Content: Evidence from Online Reviews

Samsun Knight
University of Toronto

Yakov Bart
Northeastern University

November 2023

Generative AI and User-Generated Content: Evidence from Online Reviews

Samsun Knight, University of Toronto*
Yakov Bart, Northeastern University, †

Date: November 1, 2023

Abstract

How has generative artificial intelligence affected the quality and quantity of user-generated content? Our analysis of restaurant reviews on Yelp.com and product reviews on Amazon.com shows that, contrary to prior lab evidence, use of generative AI in text generation is associated with significant declines in online content quality. These results are similar both in OLS and in two-period differences-in-differences estimation based on within-reviewer changes in AI use. We also find that use of generative AI is associated with increases in per-reviewer quantity of content, and document heterogeneity in the observed effect between expert and non-expert reviewers, with the strongest declines in quality associated with use by non-experts.

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

*samsundknight@gmail.com
†y.bart@northeastern.edu

1 Introduction

The invention and proliferation of generative artificial intelligence (AI) tools have been widely heralded as ringing in a new era in creative and textual production (Epstein et al. 2023). At the same time, researchers, industry practitioners and policymakers remain highly uncertain as to how the introduction of these tools will affect the economy (Eisfeldt et al. 2023). While recent studies have shown that generative AI can produce significant improvements, many commentators remain skeptical that such positive productivity effects are present in real-world applications, and worry that deleterious effects of at-scale "spam" creative production will dominate (Brynjolfsson, Li and Raymond 2023, Doshi and Huaser 2023, NYTimes 2023). In particular, websites that rely on user-generated content to produce their online product have reported fears that a large influx of low-quality AI-produced content will deteriorate the average user experience of their website (Burtch et al. 2023). Understanding how generative AI tends to be adopted in the real world is of utmost importance both to managers and to regulators.

In this paper, we present early evidence on the effect of real-world adoption of generative AI on content quality for 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. In both contexts, we find that use of generative AI is associated with significantly lower-quality content, as measured by the perceived usefulness (or helpfulness) of reviews. For Yelp, detected AI usage in the review production process is associated with a 15.5% standard deviation decrease in the number of "useful" votes, while for Amazon, detected AI usage is associated with a 14.8% standard deviation decrease in the number of "helpful" votes.[1] Similarly, using the two-period differences-in-differences estimation to inspect only within-reviewer changes between the pre-period (January 2022-November 2022, before ChatGPT's public release) and the post-period (December 2022-present, once ChatGPT is publicly released), 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]

[1]Effects on alternative measures, namely "funny" votes and "cool" votes, are qualitatively identical.

[2]Coefficients estimates for differences-in-differences estimation on our Amazon.com sample are

We also find evidence of significant heterogeneity in this deleterious effect of generative AI adoption across different reviewer types. Comparing reviewers with expert status badges ("Elite" reviewers on Yelp, or "Vine Voice" reviewers on Amazon) to reviewers without, we find that this negative quality effect is much larger for non-expert reviewers than for expert reviewers: non-expert reviewers drive the entirety of the observed effect for Yelp.com in our differences-in-differences specification, and have point estimates twice as large in magnitude for Amazon.com, compared to expert reviewers. While we are unable to separate whether this is due to inherent skill-based differences across reviewer types or due to different incentive structures related to review quality, this heterogeneity suggests that managers may be able to moderate this observed effect either through implementing different incentives around review production, or through careful hiring practices to avoid hiring those inclined to rely on lower-quality generative AI usage.

Lastly, we also find evidence that generative AI adoption is associated with significantly higher quantity of output for all settings. An increase of one standard deviation in adoption of generative AI is associated with a 1.2% increase in the number of Yelp reviews in the post-period and an 1.5% increase in the number of Amazon reviews in the post-period, respectively. However, these results turn into insignificant zero effects in our differences-in-differences framework, suggesting that much of these observed quantity effects in online reviews are almost entirely driven by across-reviewer variation.

The rest of this paper is structured as follows. Section 2 presents the data and descriptive statistics on the adoption of generative AI in review production on Yelp.com and Amazon.com over time. Section 3 presents our baseline econometric model and our differences-in-differences estimator. Section 4 presents our results on the effect of generative AI on output quality. Section 5 presents evidence on the heterogeneity of this effect between expert and non-expert groups. Section 6 concludes.

---

similar, although we have a far smaller matched sample of tracked reviewers in both periods for that sample, and results are not statistically significant.

2 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 forbrick-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.[3] The data also contain markers for 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 the full universe of posted Amazon.com

[3]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.

reviews for 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 present. 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.[4]

Detection of GPT in review writing

We detect the use of generative AI in the writing of the Yelp reviews using ZeroGPT, an online service for detecting the presence of generative AI-generated language in text. This service categorizes Yelp 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. To 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.

After dropping the 22.2% reviews that do not contain enough text for ZeroGPT classification, our Yelp data contain approximately 12.5 reviews, on average, for each restaurant in our sample, for a baseline Yelp dataset of 36,428 scored reviews. Our Amazon data contain 122,927 reviews with enough text for ZeroGPT classification. For outcome measures, we focus on the perceived review quality, measured in terms of the number of "useful" votes or "helpful" votes that a review received from other

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

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 et al. 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 Supplemental Appendix Table A1. 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.

Observed Patterns of Generative 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.

Over Time

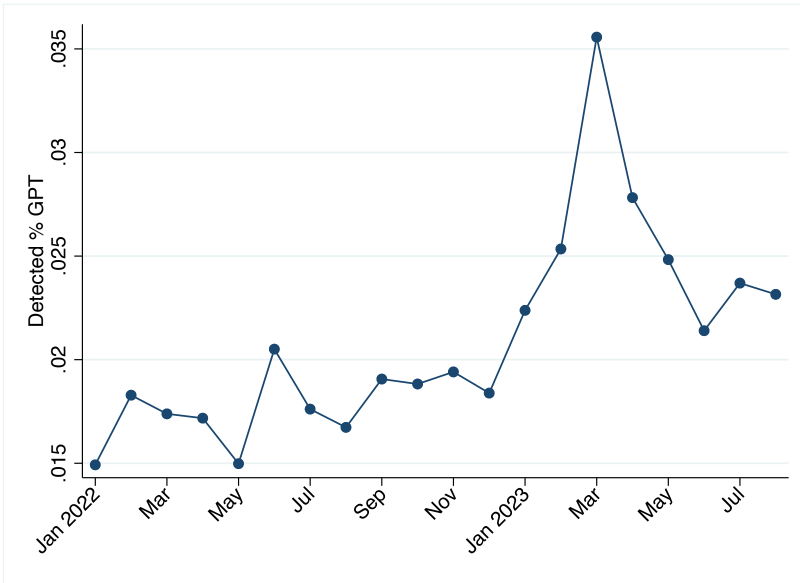
For both samples, we detect a statistically significant increase in detected generative AI adoption in review production after the public release of ChatGPT[5] that is also visually salient in the binscatters of detected AI usage rates across 2022 and the first 9 months of 2023, as presented in Figure 1. In the case of Yelp, AI adoption spikes in 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.

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 either some reviewers were using earlier versions of LLMs to produce reviews, or that some of these are false positives from the ZeroGPT detection software. In light of this, in order to ensure that our results are not driven by false positive rates in GPT

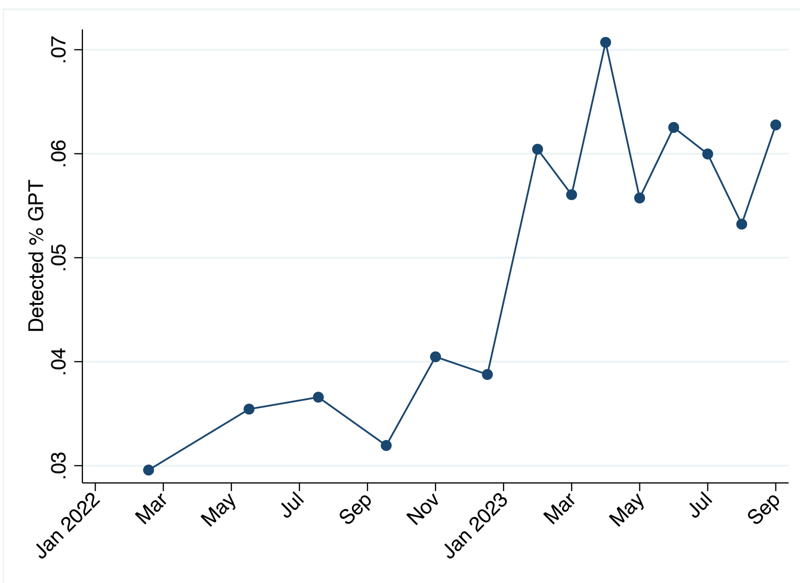
[5]Regressing an indicator for the post-release period on a dummy for whether AI was detected in the writing of the review, for each respective sample we find a highly significant increase: $\beta_{Amazon} = 0.019$, $t_{122984} = 15.8$, $p < 0.001$, and $\beta_{Yelp} = 0.007$, $t_{79235} = 6.88$, $p < 0.001$.

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 79,237 reviews gathered from Yelp.com for the universe of restaurants in San Francisco, and 122,927 reviews gathered from Amazon.com for the "Camera, Photo and Video" category.

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 ChatGPT adoption's effect on review production.

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 Supplemental Appendix Figure A1. This suggests that GPT-written reviews are not exclusively being used to "flood" either 1-star or 5-star reviews, at least in our present 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 Supplemental Appendix Figures A2 and A3. 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 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 generative AI adoption on perceived review quality in the pre-ChatGPT-release period of January 2022 through November 2022, and $\beta_{GPT-post}$ measures the effect of generative AI adoption on perceived review quality in the post-ChatGPT-release period of December 2022 and onwards.

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

With this specification, we capture the effect of generative AI adoption 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 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.[6] We then estimate the effects on quantity with separate coefficients for pre- and post-release effects of detected AI use in writing, as before:

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

Finally, to investigate results using the most extensive set of controls available to us, we further implement a specification that estimates the effect of generative AI adoption prior to ChatGPT's release and after in a two-period event-study design, to control for possible differing selection into generative AI use:

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

In this specification, the estimated $\beta_{GPT-post}$ is identified from the effect of generative AI adoption 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.

[6]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.

4 Empirical Results

4.1 Effect of Generative AI on Output Quality

We first examine the impact of 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.[7]

Results are presented in Tables 1 and 2. We find that across the board, generative AI use in the post-ChatGPT release period is associated with lower-quality production, with effects ranging from 10.4% to 15.5% 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 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.

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 only becomes stronger and more 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 reviews, presented in Supplemental Appendix Tables A2 and A3. 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, we find small and statistically null results for the effect of detected generative AI use in the pre-period.

[7] Note that we standardize outcome variables to express effects in terms of standard deviation shifts in perceived review quality.

Table 1: Effect of Detected AI Use on Yelp.com Review "Useful" Votes

$\beta_{GPT-pre}$	-0.011	-0.015
	(0.038)	(0.036)
$\beta_{GPT-post}$	-0.104**	-0.155***
	(0.032)	(0.032)
Controls		X
N	79237	79237

Table 2: Effect of Detected AI Use on Amazon.com Review "Helpful" Votes

$\beta_{GPT-pre}$	0.166***	0.080***
	(0.019)	(0.018)
$\beta_{GPT-post}$	-0.148***	-0.148***
	(0.018)	(0.017)
Controls		X
N	122927	122927

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 category dummy, and for Yelp restaurants, type category dummy. Outcome variables standardized to express effects in terms of standard deviation shifts in quality. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

4.1.1 Heterogeneity Across User Types

While the estimated negative relationship between AI adoption in review generation goes against some of the findings of earlier studies on generative AI adoption, this earlier work also emphasizes that meaningful heterogeneity exists in productivity effects across different worker types, suggesting that perhaps this negative relationship is not representative of effects for all workers.

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 "expert" reviewers below, and reviewers that do not belong to either category as "non-expert" reviewers.) In each 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 say whether any observed differences between them is due to worker type or due to the incentive structures of these programs.

Results are presented in Supplemental Appendix Tables 4 and 5. Using the same specification as above, stratified by reviewer type, we find that our measured effect on quality manifests for both expert and non-expert reviewers: in the case of Yelp.com, we find a slightly stronger negative association for expert reviewers in terms of point estimates, while in the case of Amazon.com, we find a slightly stronger negative association for low-skill reviewers. In all cases, we find a significant negative association between post-ChatGPT release generative AI usage and review reported quality that is not confounded by any pre-period effects.

4.1.2 Differences-in-Differences Estimates

While this evidence is strongly suggestive of a negative relationship between generative AI adoption 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. If worse reviewers are more likely to adopt generative AI, lower-quality reviews could be associated with generative AI use outside of a causal relationship between generative AI and output quality. Moreover, the degree of this selection effect may differ across groups, similarly confounding our comparisons across expert and non-expert groups.

To control for this and other possible confounders, we collapse our data into reviewer-level averages from January 2022 to November 2022 and December 2022 to present, in order to assess the effect of generative AI writing on a given reviewer using a 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 changes within an identified reviewer rather than spurious selection effects or other reviewer-specific or period-specific 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 3 and 4. For the Yelp.com sample, we uncover a strongly significant negative effect of 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 sample, our smaller sample does not have enough precision to estimate statistically signifi-

Table 3: 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 4: Effect of Detected AI Use on Amazon.com Review "Helpful" Votes: Differences-in-Differences

$\beta_{GPT-pre}$	-0.027	-0.028
	(0.016)	(0.016)
$\beta_{GPT-post}$	-0.025	-0.034
	(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 category dummy, and for Yelp restaurants, type category dummy. Outcome variables and covariates standardized to express effects in terms of standard deviation shifts. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

cant effects, but the point estimates suggest that a standard deviation increase in reviewer AI usage in the post-period leads to a 3.4% 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 expert and non-expert reviewers, as in section 4.1.1, to estimate how this effect differs across reviewer types. Results are presented in Tables 5 and 6. Here we see a much clearer distinction in the effects of generative AI adoption across reviewer types, as compared to the naive OLS: the negative relationship with quality is consistently much larger for non-expert reviewers. For Yelp.com reviews, the overall observed effect appears to be entirely driven by non-expert reviewers, while for the much smaller sample of Amazon.com reviews, again both point estimates are insignificant but suggest that the negative effect for non-expert reviewers is twice as large.

Notably, the distinction between the effects across these groups is far clearer and more consistent in these tables than in the observational specification of Supplemental Appendix Tables A4 and A5. This suggests that not only does the effect differ across these groups, but also selection patterns may differ as well: observed effects 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 be attenuated by *positive* selection into ChatGPT use, where the most-skilled non-expert reviewers choose to adopt generative AI in their review production. Results for Amazon.com are too noisy to make clean inferences on the plausible direction of selection effects, if any.

Overall, it appears that workers are incorporating generative AI into their production processes as a time-saving device, but are, on average, neglecting to exploit opportunities for improving output quality, documented in prior research focusing on lab studies. While the average quality in other settings may be higher, in these real-world contexts, non-expert online reviewers do not tend to leverage the technology without sacrificing a significant degree of quality in the process.

Table 5: Effect of Detected AI Use on
Yelp.com Review "Useful" Votes:
Diff-in-Diff, "Elite" vs. non-"Elite"

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

Table 6: Effect of Detected AI Use on
Amazon.com Review "Helpful" Votes:
Diff-in-Diff, "Vine Voice" vs. non-"Vine Voice"

	"Vine Voice"	Non-"Vine Voice"
$\beta_{GPT-pre}$	-0.015	-0.038
	(0.012)	(0.031)
$\beta_{GPT-post}$	-0.028	-0.055
	(0.023)	(0.049)
δ_j	X	X
γ_t	X	X
Controls	X	X
N	968	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 category dummy, and for Yelp restaurants, type category dummy. Outcome variables and covariates standardized to express effects in terms of standard deviation shifts. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

4.2 Effect of Generative AI on Output Quantity

Next, we examine the effects of generative AI adoption 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.

Results are presented in Supplemental Appendix Tables A6 and A7. 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.

4.2.1 Heterogeneity Across User Types

To investigate how this effect may be moderated by reviewer category and incentive structures, we stratify the sample between expert and non-expert reviewers based on each respective website's 'expert reviewer badge' and estimate our baseline specification for both expert and non-expert reviewer categories.[8] Results are presented in Supplemental Appendix Tables A8 and A9. 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

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

fairly broad-based across groups. For Amazon.com reviews, point estimates are quite similar across expert and non-expert reviewers, while for Yelp.com reviews, quantity effects appear to be much more pronounced for expert reviewers.

4.2.2 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 Supplemental Appendix Tables A10 and A11. Overall, we find no significant effects of generative AI adoption 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.

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 (Huang, Grady and GPT-3 2022).

This study is the first, to the best of our knowledge, that provides a rigorous evidence that these fears are well-founded. 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 generative AI use significantly reduces the average quality of reviewer output when adopted into the review production process. 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 generative AI adoption leads to a 6.5% decrease in perceived Yelp review quality, as measured by the number of "useful" votes that the review receives.

We also find suggestive evidence that these effects differ across reviewer categories, and that employers may not be helpless to prevent such use: for reviewers who are evaluated on the quality of their reviews (with "Elite" badge), generative AI use does not harm content quality. The negative effects are concentrated, rather, among reviewers with little to no incentive to produce high-quality reviews. For these reviewers only does generative AI use lead to sharp and significant declines in perceived review quality.

All the same, in the laissez-faire content generation environments that we inspect, responsible usage does not appear to win out. With this work, we hope to sound an early warning bell at this possible worse-off equilibrium, and to encourage future research that might explore pragmatic approaches to capture and leverage the potential benefits of generative AI without such drawbacks. Future research may also extend this analysis to investigate the effect of generative AI use on content quality in other contexts, and to measure the effect of generative AI use on other aspects of production.

Generative AI will transform the world. But right now it needs guidance. Present 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, better-off society.

References

Brynjolfsson E, Li D, Raymond LR (2023) Generative ai at work. Working Paper 31161, National Bureau of Economic Research, URL http://dx.doi.org/10.3386/w31161.

Burtch G, Lee D, Chen Z (2023) The consequences of generative ai for ugc and online community engagement. Available at SSRN 4521754.

Doshi AR, Hauser O (2023) Generative artificial intelligence enhances creativity but reduces the diversity of novel content Available at SSRN: https://ssrn.com/abstract=4535536 or http://dx.doi.org/10.2139/ssrn.4535536.

Eisfeldt AL, Schubert G, Zhang MB (2023) Generative ai and firm values (w31222).

Epstein Z, Hertzmann A, Investigators of Human Creativity, Akten M, Farid H, Fjeld M, Frank MR, Groh M, Herman L, Leach N, Mahari R (2023) Art and the science of generative ai. *Science* 380(6650):1110–1111.

Huang S, Grady P, GPT-3 (2022) Generative ai: A creative new world. https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/, accessed: November 1, 2023.

Hutto C, Gilbert E (2014) Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 216–225.

Roose K (2023) GPT-4 Is Exciting and Scary. *The New York Times* ISSN 0362-4331, URL https://www.nytimes.com/2023/03/15/technology/gpt-4-artificial-intelligence-openai.html.

Table A1: 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	.	(.)	6.45	(32.30)
Star rating	4.15	(1.21)	4.29	(1.06)
Valence	0.22	(0.11)	0.16	(0.10)
Wordcount	108.50	(88.13)	166.67	(196.50)
Observations		79237		122986

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.

Table A2: Effect of ChatGPT on Review "Funny"-ness: Yelp.com

$\beta_{GPT-pre}$	-0.035 (0.038)	-0.036 (0.037)
$\beta_{GPT-post}$	-0.080* (0.032)	-0.124*** (0.032)
Controls		X
N	79237	79237

Table A3: Effect of ChatGPT on Review "Cool"-ness: Yelp.com

$\beta_{GPT-pre}$	-0.023 (0.038)	-0.037 (0.037)
$\beta_{GPT-post}$	-0.085** (0.032)	-0.139*** (0.033)
Controls		X
N	79237	79237

Notes: Effect of adoption of ChatGPT 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 category dummy, and for Yelp restaurants, type category dummy. Outcome variables standardized to express effects in terms of standard deviation shifts in quality. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

Table A4: Effect of Detected AI Use on Yelp.com Review "Useful" Votes: "Elite" vs. non-"Elite"

	"Elite"	Non-"Elite"
$\beta_{GPT-pre}$	-0.028	0.047
	(0.060)	(0.041)
$\beta_{GPT-post}$	-0.198***	-0.076*
	(0.048)	(0.039)
Controls	X	X
N	30735	48502

Table A5: Effect of Detected AI Use on Amazon.com Review "Helpful" Votes: "Vine Voice" vs. non-"Vine Voice"

	"Vine Voice"	Non-"Vine Voice"
$\beta_{GPT-pre}$	0.079	0.077***
	(0.047)	(0.019)
$\beta_{GPT-post}$	-0.073**	-0.137***
	(0.027)	(0.020)
Controls	X	X
N	15463	107464

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 category dummy, and for Yelp restaurants, type category dummy. Outcome variables standardized (separately for each subsample) to express effects in terms of standard deviation shifts in quality. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

Table A6: 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 A7: Effect of Detected AI Use on
Log Number of Amazon Reviews

$\beta_{GPT-pre}$	-0.007***	-0.009***
	(0.002)	(0.002)
$\beta_{GPT-post}$	0.022***	0.015***
	(0.002)	(0.002)
Controls		X
N	52237	52237

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 star rating of review, wordcount of review, average valence of review, price category dummy, and for Yelp restaurants, type category dummy. Outcome variables standardized to express effects in terms of standard deviation shifts in quality. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

Table A8: Effect of Detected AI Use on
Log Number of Yelp Reviews:
"Elite" vs. non-"Elite"

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

Table A9: Effect of Detected AI Use on
Log Number of Amazon Reviews:
"Vine Voice" vs. non-"Vine Voice"

	"Vine Voice"	Non-"Vine Voice"
$\beta_{GPT-pre}$	-0.023**	-0.007***
	(0.007)	(0.002)
$\beta_{GPT-post}$	0.013	0.012***
	(0.007)	(0.002)
Controls	X	X
N	6208	46029

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 star rating of review, wordcount of review, average valence of review, price category dummy, and for Yelp restaurants, type category dummy. Outcome variables standardized to express effects in terms of standard deviation shifts in quality. * = 5% significance, ** = 1% significance, *** = 0.1% significance.

Table A10: Effect of Detected AI Use on
Log Number of Yelp Reviews:
Differences-in-Differences

$\beta_{GPT-pre}$	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

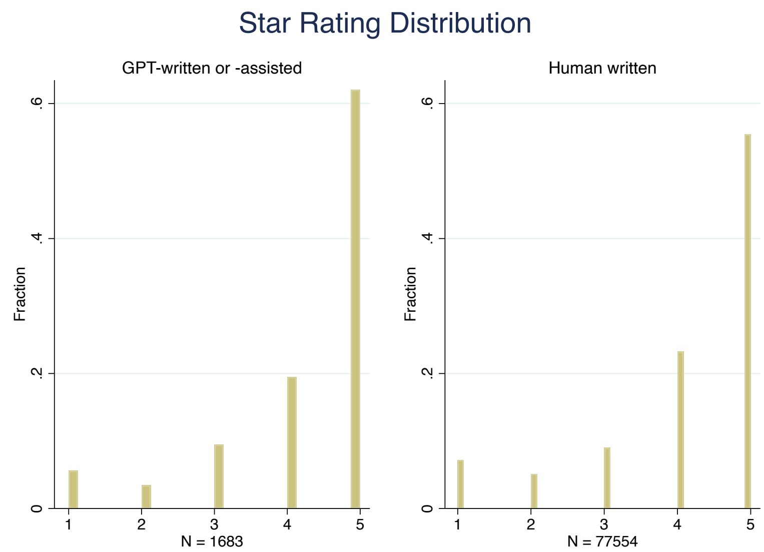
Table A11: Effect of Detected AI Use on
Log Number of Amazon Reviews:
Differences-in-Differences

$\beta_{GPT-pre}$	-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

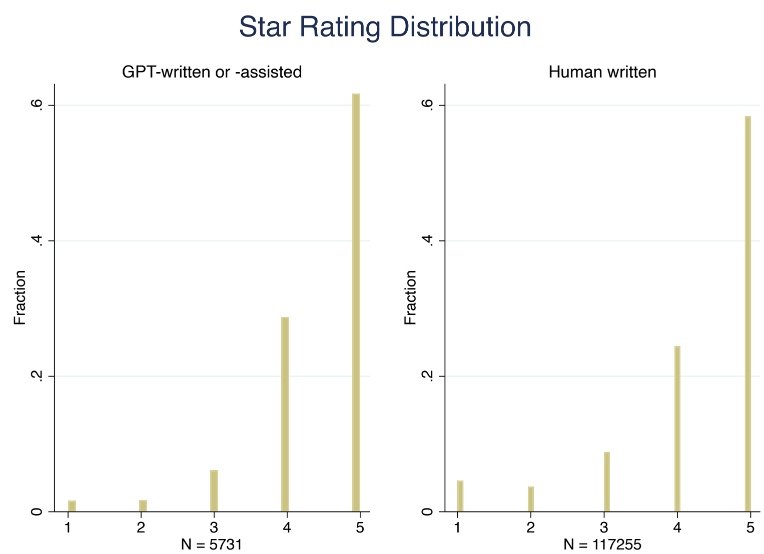
Notes: Difference-in-differences estimates of the effect of adoption of ChatGPT in review production on review quantity produced per reviewer, per 11 months. Sample based on Yelp reviews for restaurants in San Francisco area and reviews for products in "Camera, Photo and Video" category of Amazon, with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

Figure A1: Star Rating Distribution:
Human versus Detected AI

(a) Yelp.com



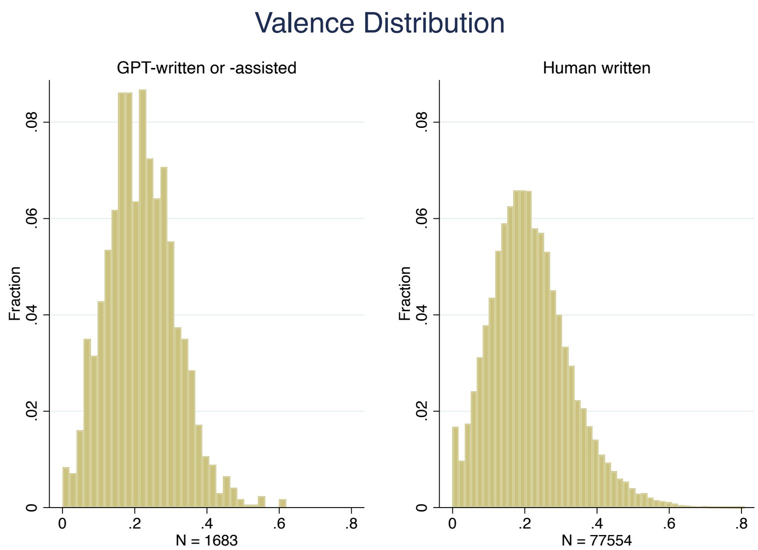
(b) Amazon.com



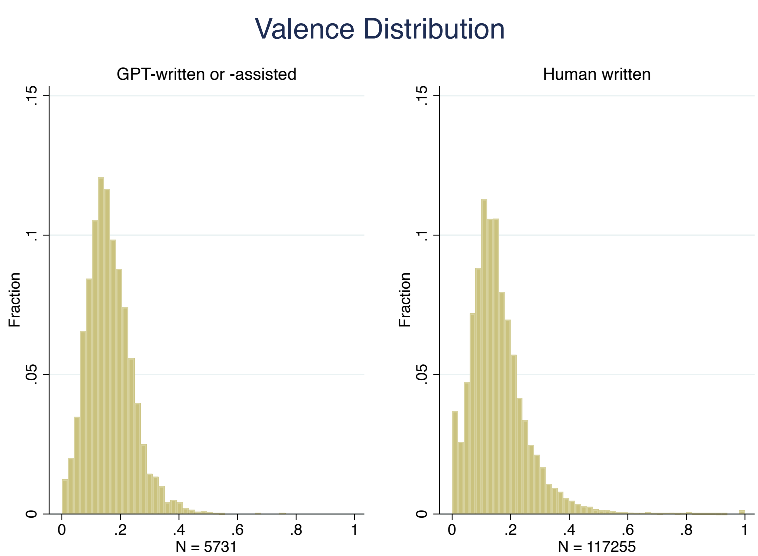
Notes: Star rating distribution of reviews, stratified by human-generated or generative AI-assisted. Y-axis of each panel shows the fraction of reviews at each star rating, while X-axis shows the associated star rating. Use of generative AI scored using ZeroGPT.com. Sample based on 79,237 reviews gathered from Yelp.com for the universe of restaurants in San Francisco, and 122,927 reviews gathered from Amazon.com for the "Camera, Photo and Video" category.

Figure A2: Valence Distribution:
Human versus AI

(a) Yelp.com



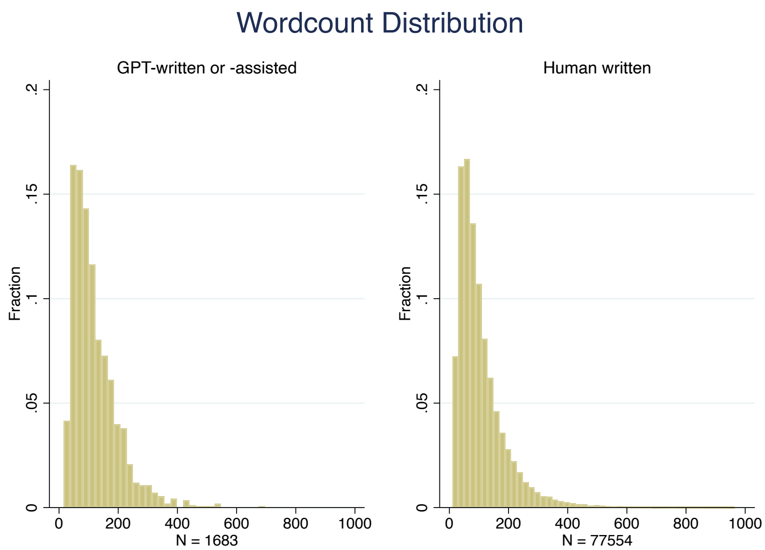
(b) Amazon.com



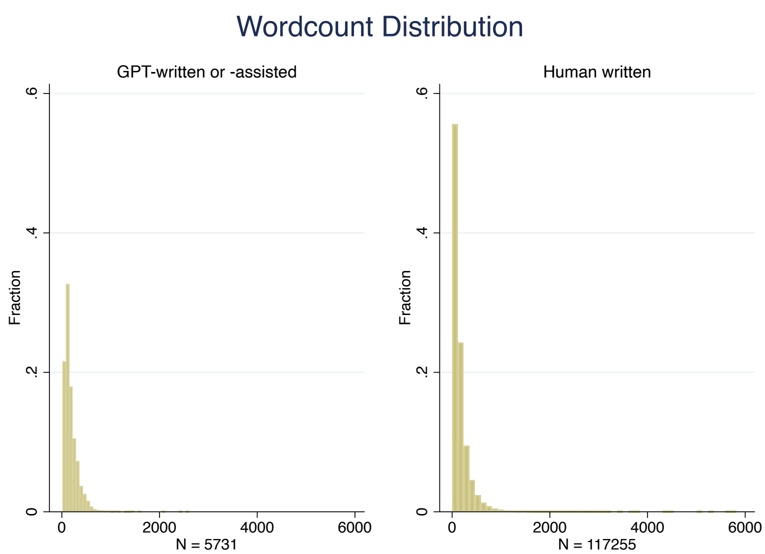
Notes: Valence distribution of reviews, stratified by human-generated or generative AI-assisted. Y-axis of each panel shows the fraction of reviews at each valence, while X-axis shows the associated valence. Use of generative AI scored using ZeroGPT.com. Sample based on 79,237 reviews gathered from Yelp.com for the universe of restaurants in San Francisco, and 122,927 reviews gathered from Amazon.com for the "Camera, Photo and Video" category.

Figure A3: Wordcount Distribution:
Human versus AI

(a) Yelp.com



(b) Amazon.com



Notes: Wordcount distribution of reviews, stratified by human-generated or generative AI-assisted. 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 79,237 reviews gathered from Yelp.com for the universe of restaurants in San Francisco, and 122,927 reviews gathered from Amazon.com for the "Camera, Photo and Video" category.