AI Summary Overrepresents Fake Reviews: Evidence from Amazon

(Preliminary Draft)

Sihan Zhai* and Andrew T. Ching†

*Harvard Business School
†Carey Business School, Johns Hopkins University

December 28, 2025

Abstract

AI summary has been widely deployed to distill information from large volumes of reviews. In this research, we argue that there is an unintended consequence of AI summary – it tends to overrepresent fake reviews. Our key insights are (i) AI summary algorithms focus on extracting sentiments of common themes (i.e., keywords) from reviews, (ii) fake reviews are more similar to each other, leading to more common themes, and further a stronger influence on the overall sentiment of each common theme. To provide empirical support for our argument, we study AI summary on Amazon. We use Amazon-sold products as a proxy for products less likely to have fake reviews, and products with low review credibility grades on RateBud as a proxy for products more likely to have fake reviews, and we also identify fake reviews documented in existing literature. We first confirm that fake reviews are significantly longer, more positive (even than authentic reviews with the same rating), and more similar to each other. We also find that, for products more likely to contain fake reviews, extracted keywords are mentioned by a larger number of reviews, indicating greater concentration of reviews around common themes. We conjecture these mentions are dominated by fake reviews. Given that fake reviews are more positive, we further investigate if keyword's sentiment tends to be more positive for products more likely with fake reviews – we indeed find robust evidence to support this empirical implication. Moreover, we find that compared with other products, the overall sentiments of AI summary of products with more fake reviews are relatively more positive than average ratings. Finally, we study the market-distortion effect. We find evidence that the bias of AI summary improves sales of products sold by review manipulators.

Keywords: AI Summary, AI Bias, Fake Reviews, Digital Platforms, Misinformation

1 Introduction

Information asymmetry is a persistent problem on digital platforms: sellers know more about the quality of products than consumers. Online reviews can reduce this asymmetry, but they are polluted by paid, inflated reviews posted to mislead consumers. As a result, fake reviews weaken the informational value of review systems and can harm consumer welfare—prompting sustained attention from both the media and regulators. Major media outlets—such as the Wall Street Journal and the Associated Press—repeatedly warn that fake reviews are widespread,[1,2] and that they can fool even experienced shoppers.[3] This concern has prompted regulators to take action: the Federal Trade Commission (FTC) of the US[4] and the Parliament of the UK[5] are both trying to combat fake reviews.

Digital platforms, such as Amazon, Walmart, Best Buy, Yelp, Apple App Store, and Google Play Store, have recently deployed AI summary of reviews. According to Vaughn Schermerhorn, the director of community shopping at Amazon, AI summary was introduced to help consumers more easily grasp common themes from the vast volume of reviews on the platform.[6] AI summary typically presents two user-facing parts: (i) a list of keywords with their associated sentiment labels, and (ii) a short summary paragraph. The two parts are generated in the following procedure. The keyword-extraction models in AI summary first extract keywords commonly mentioned across reviews and sentiment-analysis models assess the average sentiment of reviews that mention each keyword. The short paragraph is composed by a language model, based on the extracted keywords and sentiments. Crucially, reviews are not treated equally in this process; "common themes" are extracted as keywords, while more unique opinions, which cannot be classified into any themes, tend to be neglected.

We argue that AI summary may unintentionally amplify fake reviews. Fake reviews are likely written without idiosyncratic first-hand experience, and therefore emphasize generic merits and common themes. Moreover, paid fake reviewers are asked to write longer reviews, increasing textual overlap across fake reviews and making them more likely to concentrate on shared themes. Therefore, AI summary designed to extract "common themes" may overrepresent fake reviews. We test whether more fake reviews mention common themes, and are therefore more represented in the analysis of sentiments corresponding to keywords. Moreover, we hypothesize that fake

reviews are more positive than authentic reviews because they are paid by review manipulators. Therefore, AI summary may favor review manipulators by disproportionately highlighting their overly positive fake reviews in the sentiments of AI summary. We further test whether AI summary algorithms generate more positive sentiments on both keywords and summary paragraphs of review manipulators. Last but not least, we test whether AI summary distorts market outcomes by directing more sales to review manipulators.

To test these predictions, we assemble a dataset from multiple sources, including Amazon, Keepa, RateBud, and the existing literature. We use Amazon-sold products as a proxy for products less likely to have fake reviews, because Amazon almost never engages in shady review manipulation (He, Hollenbeck, and Proserpio, 2022), and use products with low review-credibility grades on RateBud[7] as a proxy for products more likely to have fake reviews, and we also leverage fake reviews documented in the existing literature (He et al., 2022; Feldman, Tosyali, and Overgoor, 2025).

We first test whether fake reviews contribute more to the sentiments of AI summary. By analyzing the textual features of fake reviews in the literature, we confirm that fake reviews are more similar to each other, more positive, and longer than even authentic positive reviews. In addition, when we apply the common methods used in keyword-extraction algorithms to our data, we find that fake reviews are more clustered in the space of embeddings. Next, we download keywords from Amazon. We find that a greater portion of reviews align with the keywords extracted by Amazon, so that they are input into sentiment analysis and consequently represented in the sentiments of AI summary.

Second, we examine whether fake review products receive systematically more positive AI summary. We download sentiments corresponding to keywords and analyze sentiments expressed in the summary paragraphs with a general-purpose model OpenAI GPT-4.1 and a finetuned BERT specialized in sentiment analysis of reviews. We find that after controlling for the distribution of ratings and categories of products, compared with other products on the same search result page, products more (less) likely to have fake reviews have more (less) positive sentiments associated with keywords and AI summary paragraphs. Moreover, we find that fake reviews boost AI summary more than average ratings.

For the above two results, we run falsification tests. Sellers generally have to cover the costs of

buying fake reviews, so the costs of manipulating reviews of expensive products are higher. We find that the above-mentioned phenomena disappear among the subset of expensive products, for which review manipulation is more costly. There are a variety of differences between subgroups. This finding suggests that the results discovered between products more and less likely to have fake reviews are likely due to the difference in the number of fake reviews rather than other differences between subgroups.

Finally, we turn to analyze whether after the introduction of AI summary, the sales of review manipulators increased. We find that compared with other products, following the introduction of AI summary, sales ranks of Amazon-sold (thus less manipulable) products fell by around 10%, while those with likely fake reviews rose by around 20%.

This paper contributes to three streams of literature. We document a new type of AI bias, which is an unintended consequence due to the mechanistic nature of keyword extraction in AI summary algorithms and the textual features of fake reviews. We contribute to the literature on fake reviews by showing that AI summary can make the sentiments of fake reviews more salient and increase the sales of products with fake reviews, and it might negatively impact consumer welfare. We contribute to the literature on the transmission of misinformation by showing that AI can function as an amplifier of misinformation, which is fake reviews in our context.

The findings of the paper are important to policymakers. If platforms continue to deploy AI summary of reviews without correcting the AI bias documented here, it may lead to welfare loss for consumers. This research is also relevant to platforms that care about long-term customer trust in their services, because the current AI summary algorithms might drive consumers to make suboptimal decisions, and can further make customers leave the platform. We also hope this research will inspire AI developers to invent better algorithms to summarize information contaminated by misinformation, because simply improving the accuracy of current algorithms only makes the bias even worse.

The remainder of the paper is organized as follows. Section 2 reviews the literature and elaborates on our contributions to the literature. Section 3 introduces the business settings by synthesizing information from media, industry reports and academic research, which naturally leads to hypotheses to test. Section 4 introduces data sources and our proxies for products with fake reviews. Section 5 presents empirical results. Section 6 concludes with the implications of the

findings.

2 Literature Review

Our work intersects three streams of literature, i.e., AI bias, fake reviews, and spread of misinformation, by studying how AI summary is designed to be biased towards review manipulators. The literature on AI bias and algorithmic bias is emerging and is attracting the attention of scholars in various fields. Cowgill and Tucker (2019) provide a comprehensive review of the literature on algorithmic bias. In this paper, they distinguish algorithmic bias caused by biased objectives from algorithmic bias caused by biased predictions. They highlight the particular challenges in dealing with algorithmic bias caused by biased objectives. Solving algorithmic bias caused by biased objectives is a governance problem, which involves leadership, ethics, and oversights, rather than pure technical tweaks. A classical example of algorithmic bias caused by biased objectives is filter bubble. Researchers in different fields (Levy, 2021; Nyhan et al., 2023; González-Bailón et al., 2023) have recently confirmed that recommendation algorithms deployed by social media platforms expose users to more like-minded content that is consistent with users' own political beliefs. This is because the objective of recommendation algorithms is to maximize engagement, while consumers are more likely to engage with like-minded content (Robertson et al., 2023). Researchers are concerned that filter bubble may lead to political polarization (Levy, 2021). Other examples of objective-induced algorithmic bias include Obermeyer and Mullainathan (2019), in which they find racial disparities in a commercial health risk prediction algorithm that mistakenly optimized for healthcare cost rather than actual health. For another example, Lambrecht and Tucker (2019) find that algorithms that distribute advertisements favor cheaper exposures to consumers. However, exposures to male consumers are cheaper because commercial advertisers compete intensively for female consumers. Consequently, advertisements advocating STEM job information are more likely to reach male audience. Our research contributes to this stream of literature by documenting a new example of AI bias induced by biased objectives. We show that such AI summary powered by language models favors fake reviews because they are unintentionally designed to seek textual features typically possessed by fake reviews. As AI summary of reviews becomes ubiquitous as a result of the remarkable progress of language models, the AI bias that we document becomes

increasingly relevant.

We also contribute to the literature on fake reviews. Online reviews have been shown to influence the behaviors of both consumers (Chevalier and Mayzlin, 2006; Zhu and Zhang, 2010; Sun, 2012; Lu et al., 2022; Wang, Tong, and Dong, 2025) and sellers (Farronato and Zervas, 2022), and consequently influence consumer welfare (Wu et al., 2015). Therefore, the manipulation of online reviews has raised widespread concerns. Yet, the inherently covert nature of review manipulation hinders direct observation of fake reviews and, consequently, empirical analysis. Mayzlin, Dover, and Chevalier (2014), who creatively exploit economic incentives to manipulate reviews, are among the first to show convincing evidence of fake reviews. Luca and Zervas (2016) further investigate the economic incentives underlying review manipulation by studying the reviews filtered and then removed by Yelp. He, Hollenbeck, and Proserpio (2022) represent a significant step forward in the literature by directly observing sellers who buy fake reviews in Facebook groups, and therefore more confidently identifying fake reviews on Amazon. Not only do they provide interesting details on how the market for fake reviews works and the textual features of fake reviews, they also shed light on whether review manipulation hurts consumer welfare in the theoretical literature (Dellarocas, 2006; Mayzlin, 2006). Specifically, they find that after sellers bought fake reviews, the negative reviews written by likely cheated consumers increased, suggesting that consumer welfare is hurt by review manipulation. Gandhi, Hollenbeck, and Li (2024) further study the equilibrium effects and welfare implications of fake reviews using a structural model. Our paper contributes to this stream of literature by suggesting that the negative welfare impact of fake reviews might be reinforced by AI summary in the AI era.

Our paper is related to the literature on the transmission of misinformation. Vosoughi, Roy, and Aral (2018) provide strong empirical evidence that misinformation diffuses significantly faster, farther, deeper, and more broadly than authentic information. This finding challenges the idea of wisdom of the crowd, by showing that a huge group of people are behaving irrationally and fail to correct each other. Researchers are interested in the underlying reason why misinformation is more likely to be transmitted. In the theoretical literature, Acemoglu, Ozdaglar, and ParandehGheibi (2010) and Acemoglu et al. (2013) find persuasive agents who are trying to influence others or are not interested in updating their own beliefs in the social network can facilitate the transmission of misinformation and help false beliefs survive. Mostagir, Ozdaglar, and Siderius (2022) identify

social network structures that are susceptible to information manipulation. The existing empirical studies also find evidence that robots (Shao et al., 2018), the tendency of human beings to engage more with misinformation (Vosoughi, Roy, and Aral, 2018), the failure of human beings to consider accuracy of information (Pennycook et al., 2021), overconfidence in the ability to spot misinformation (Lyons et al., 2021), political polarization (Zhu and Pechmann, 2024), and echo chambers (Vicario et al., 2016; Törnberg, 2018) can facilitate spread of misinformation. Our paper contributes to this stream of literature by showing that AI can also amplify the voice of false information, which is fake reviews in our context, on digital platforms.

3 Settings and Hypotheses

In this section, we discuss the institutional background about AI summary algorithms and fake reviews. We then derive hypotheses to be tested later.

3.1 AI Summary of Reviews

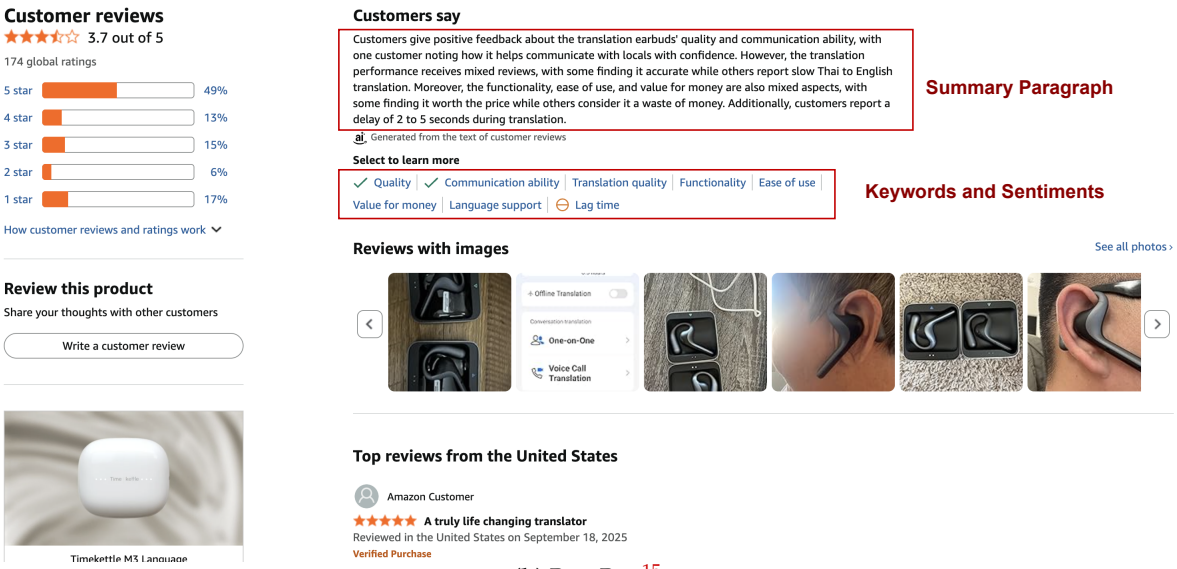
With the development of language models, AI summary of reviews is becoming a common business practice on digital platforms. AI summary of reviews is machine-produced and language-model–based condensations of user-generated reviews that aim to mitigate information overload and support consumer decision-making on digital platforms. Most digital platforms, including Amazon,[8] Walmart,[9] Best Buy,[10] Yelp,[11] Apple App Store[12] and Google Play Store[13] have adopted AI summary of reviews and they are showing AI summary in a prominent location above all user-generated reviews.

Interfaces of AI summary on the websites of Amazon and Best Buy are shown in Figure 1. AI summary typically includes two user-facing parts, a list of keywords with corresponding sentiments, and a summary paragraph. Some platforms have both (e.g., Amazon and Best Buy), some only have keywords with sentiments (e.g., Yelp), and some only have summary paragraphs (e.g., Apple App Store). When both keywords with sentiments and summary paragraphs are present, summary paragraphs typically connect keywords into a human-understandable summary paragraph so that the two user-facing parts are consistent with each other.

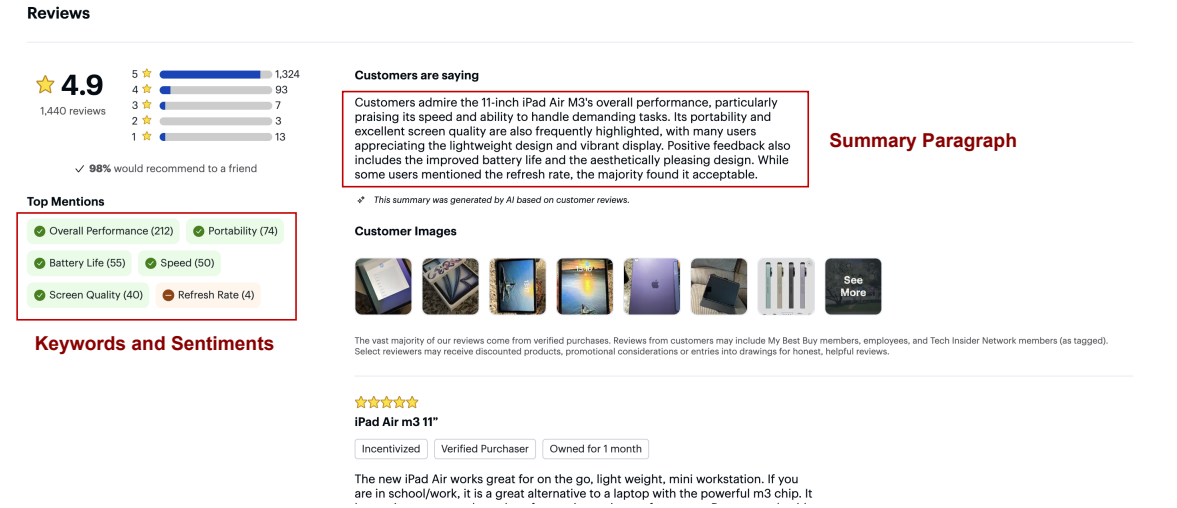
In this paper, we focus on the major player, Amazon. More than half (56%) of consumers in the

Figure 1: User Interfaces of AI Summary

(a) Amazon[14]



(b) Best Buy[15]



US start their shopping searches on Amazon.[16] Amazon introduced AI summary of reviews on August 14, 2023.[6] As for the business goal of AI summary, Vaughn Schermerhorn, the director of community shopping at Amazon, claimed that they "want to make it even easier for customers to understand the common themes across reviews", and that they are able to do so thanks to the advancement in generative AI. In terms of what AI summary is like, he mentioned that it "provides a short paragraph right on the product detail page that highlights the product features and customer sentiment frequently mentioned across written reviews to help customers determine at a glance whether a product is right for them."

A document from their internal AI team describes the basic features of the algorithm.[17] Ac-

cording to Amazon engineers, AI summary "assigns sentiment analysis and keyword extraction to traditional ML while using optimized SLMs (small language models) for complex text generation tasks". In other words, Amazon first feeds user-generated reviews into specialized traditional machine learning models to extract keywords commonly mentioned across reviews. In this process, reviews that do not match any keywords are excluded from AI summary. On Amazon, if we click on keywords, we can see each keyword is mentioned by how many reviews. One review can align with multiple keywords simultaneously.[18] Then, Amazon inputs reviews that mention each keyword into sentiment-analysis models to analyze their sentiments, and displays the output sentiments using a sign next to keywords. Finally, Amazon uses language models to generate the summary paragraphs.[19] We should highlight one key difference of AI summary from traditional methods of summarization, such as average ratings. Instead of extracting information equally from all reviews, the AI summary algorithm only considers sentiments of reviews that mention the extracted keywords, while ignoring the remaining reviews.

There is very little existing research studying AI summary. To our knowledge, the very few existing papers are still work-in-progress. Wang, Tong, and Dong (2025) find that the introduction of AI summary increases purchase rates and shifts users from reading more reviews to exploring more listings. Su et al. (2025) study the impact of AI summary on the diversity of following user-generated reviews. To the best of our knowledge, no existing research has studied whether misinformation is overrepresented by AI summary.

3.2 Fake Reviews

Amazon sellers who attempt to manipulate reviews (i.e., review manipulators) create private groups on social media platforms such as Facebook, Twitter, and Telegram to recruit fake reviewers.[20] These review manipulators offer consumers deals to write 5-star reviews for them in exchange for a full refund and, sometimes, an additional payment. Amazon itself does not engage in this type of fake review practice (He, Hollenbeck, and Proserpio, 2022), so Amazon-sold products are less likely to have fake reviews. Due to the high costs of buying fake reviews, there are very few negative fake reviews for competitors on Amazon (He, Hollenbeck, and Proserpio, 2022). Also related to the high price of fake reviews, according to Saoud Khalifah, the founder of Fakespot, the

main buyers of fake reviews are sellers of low-price products (around $15 - $40).

Research studies the textual features of fake reviews. Ott et al. (2011) find that fake reviews contain more superlatives, have more positive and fewer negative emotion terms, and include less concrete language than authentic reviews. Li et al. (2014) reinforce that fake reviews have overly highlighted sentiment. Luca and Zervas (2016) also have similar findings in the reviews filtered and then removed by Yelp. Moreover, Researchers notice that many fake reviews are highly similar to each other (e.g., Jindal and Liu, 2008; Mukherjee, Liu, and Glance, 2012; Rayana and Akoglu, 2015). Additionally, He, Hollenbeck, and Proserpio (2022) notice that fake reviews are longer than authentic reviews on Amazon.

3.3 Hypothesis Development

There is a consensus in the literature that compared with authentic reviews, fake reviews are more positive (Ott et al., 2011; Li et al., 2014; Luca and Zervas, 2016) and more similar to each other (Jindal and Liu, 2008; Mukherjee, Liu, and Glance, 2012; Rayana and Akoglu, 2015). Additionally, on Amazon, He, Hollenbeck, and Proserpio (2022) also find that fake reviews are longer. These findings are consistent with institutional backgrounds and economic incentives. When we have to write favorable comments for an item that we are not familiar with, we tend to include common and general merits, making fake reviews resemble each other. In contrast, when we have an idiosyncratic good personal experience with the products, we tend to include more heterogeneous details, exactly like strong recommendation letters in academia. For platforms with strict anti–fake-review policies, such as Amazon, sellers need to reimburse fake reviewers for purchasing the products in the vast majority of cases, which raises manipulation costs; consequently, paid reviewers are asked to write longer and more positive reviews. This leads us to Hypothesis 1a, which focuses on the textual features of fake reviews.

Hypothesis 1a (Textual Features of Fake Reviews)**.** *Compared with authentic positive reviews, fake reviews are longer, more positive (even with the same ratings) and more similar to each other.*

Two factors make it reasonable to hypothesize that fake reviews cluster around common themes. We can view the generation of reviews as sampling from a set of themes. Fake reviews and authentic reviews differ in two important ways. First, the set of themes from which fake reviews are drawn

is a limited number of general merits, while the set of themes for authentic reviews is abundant personal experiences. Second, fake reviews are longer, so more themes are drawn in fake review generation than in authentic review generation. In a word, in fake review generation, more themes are drawn from a smaller set of themes. Therefore, fake reviews are more likely to share at least one common theme with each other.

We know that AI summary is designed to summarize common themes across different user-generated reviews,[6] and that AI summary is powered by a keyword-extraction model, a sentiment-analysis model and a language model that finally composes a paragraph.[17] The process systematically overrepresents fake reviews: keyword-extraction algorithms by design search for common themes (Shapira et al., 2021; Liu and Lapata, 2019), so they extract keywords disproportionately more from fake reviews that cluster around common themes. Fake reviews are therefore more likely to mention the extracted keywords, to be input into sentiment analysis, and finally to be represented in the sentiments of keywords and AI summary paragraphs. This leads to Hypothesis 1b.

Hypothesis 1b (Overrepresentation of Fake Reviews)**.** *Compared with authentic reviews, more fake reviews mention common keywords, so they are more likely to be input into sentiment analysis and represented in sentiments of AI summary.*

Almost all fake reviews are positive (He, Hollenbeck, and Proserpio, 2022). Furthermore, fake reviews are asked to be more positive than authentic reviews, and even authentic positive reviews (as in Hypothesis 1a). The overrepresentation of the positive portion among all reviews of review manipulators can already make AI summary of manipulators more positive. The overrepresentation of such overly positive fake reviews makes AI summary of review manipulators even more positive. This leads to Hypothesis 2.

Hypothesis 2a (Bias in Sentiments of AI Summary)**.** *Products that are more likely to have fake reviews have more positive sentiments corresponding to extracted keywords and more positive AI summary paragraphs.*

Another empirical implication of the foregoing rationale is that fake reviews boost AI summary more than average ratings. The arithmetic mean of star ratings treats each review equally. However,

AI summary assigns higher weights to fake reviews with more common themes, so AI summary is much more responsive to fake reviews that are overly positive than average rating is. This leads to Hypothesis 2b.

Hypothesis 2b (Bias in Sentiments of AI Summary)**.** *Products that are more likely to have fake reviews have larger differences between sentiment (measured on the scale of ratings) of AI summary paragraphs and average ratings.*

We discuss the economic implications of Hypothesis 2. Researchers have long been aware that online reviews strongly influence sales (Chevalier and Mayzlin, 2006; Zhu and Zhang, 2010; Sun, 2012; Lu et al., 2022; Wang, Tong, and Dong, 2025). AI summary of reviews is positioned in a more salient location above all user-generated reviews. Hence, AI summary of reviews should have a stronger influence on sales and consumer decision-making. If Hypothesis 2 is true, we should observe that after the introduction of AI summary, review manipulators experience greater sales compared with non-manipulators. This leads to Hypothesis 3. Hypothesis 3 is a market distortion, in which cheaters benefit while truth-tellers suffer. This may direct consumers to purchase sub-optimal products and encourage more sellers to manipulate reviews. AI summary may increase the short-run sales volumes of the platforms (Wang, Tong, and Dong, 2025), but if it encourages manipulation and harms consumer welfare, the long-term profits of the platforms and social welfare can be harmed.

Hypothesis 3 (AI Summary Benefits Review Manipulators)**.** *After the introduction of AI summary, compared with other products, products that are more likely to have fake reviews experience greater sales.*

3.4 Empirical Strategy

When testing the hypotheses mentioned above, the most significant challenge is to identify which products are more (or less) likely to have fake reviews and which reviews are more (or less) likely to be fake. We use the three proxies detailed in 4.2 for fake reviews or fake review products. They are reviews documented by existing research (He et al., 2022; Feldman, Tosyali, and Overgoor, 2025) to be more likely to be fake, Amazon-sold products that almost never buy fake reviews, and products with low review credibility score on RateBud.[7] A concern is that there may be other

differences in addition to the number of fake reviews between the groups more or less likely to have fake reviews. We run falsification tests in Sections 5.1.3 and 5.2.3 to mitigate this concern.

When testing Hypothesis 1a, we simply compare the text features of fake reviews documented in the literature (He et al., 2022; Feldman, Tosyali, and Overgoor, 2025) with other reviews of the same product. When testing Hypothesis 1b, we run keyword extraction by ourselves on the review data, and also analyze keywords extracted by Amazon's algorithms. We compare how many reviews mention the common keywords and are therefore taken into the sentiment analysis across products more or less likely to have fake reviews.

Hypothesis 2 is more straightforward to test. We just compare the sentiments analyzed with machine learning models across products more or less likely to have fake reviews. To test Hypothesis 3, we obtain sales rank data from Keepa.[21] Then, we compare the change after the introduction of AI summary across products more or less likely to have fake reviews.

4 Data

4.1 Data Source

To test the hypotheses, we rely on four data sources. The first is Amazon. This is our main dataset. We download from Amazon AI summary paragraphs, extracted keywords with corresponding sentiments, the number of reviews mentioning each keyword, the total number of reviews, distribution of ratings (shares of 1-star, 2-star, 3-star, 4-star, and 5-star reviews), average ratings, product names, prices of products, categories of products, and whether the product is sold by Amazon. We identify the most popular search terms in the US and around the world from reports by commercial consulting firms, including Exploding Topics[22] and Glimpse.[23] Then, we search these search terms on Amazon, and record products on the first seven pages in the results.[24] We collected 16,921 products for 163 search terms in total on September 29, 2025. We report the summary statistics in Table 1. The number of reviews varies a lot across different products. The mean is much larger than the median, suggesting that the number of reviews follows a strongly right-skewed (positively skewed) distribution. There are a small share of products with many reviews. The price also follows a strongly right-skewed distribution. In contrast, the average ratings have a very small variation. Its distribution is concentrated in the range from 4.3 to 4.7. In

the data we collected, around 30% of the products were sold by Amazon. The missing average ratings in the data are mainly due to no reviews, and the missing Amazon-Sold product indicators are mainly because of no items in stock when we collected the data.

The second data source is a dataset compiled by Brett Hollenbeck from various studies, publicly available on his website.[25] It contains products predicted to have bought fake reviews by He et al. (2022) and fake reviewers predicted by Feldman, Tosyali, and Overgoor (2025). We use this dataset to analyze textual features of fake reviews. In the summary statistics reported in Table 1, most of the products are fake review purchasing products. The number of reviews is much lower than in data source 1. We have confirmed that this is a small subset of all reviews.[26] If we classify a review as a fake review if it is (i) a 5-star review, (ii) written for a review manipulating seller identified by He et al. (2022), and (iii) written by a fake reviewer identified by Feldman, Tosyali, and Overgoor (2025), a very high portion (67.71%) of them is fake.

The third data source is RateBud.[7] We collect from RateBud the list of suspicious review manipulators among products in data source 1. RateBud includes the grades for almost all the products in data source 1. RateBud comprehensively grades the overall credibility of reviews of products on Amazon according to product consistency (consistency between reviews and product), seller reputation (seller credibility with brand recognition), review distribution (whether distribution of ratings is natural), reviewer credibility (trustworthiness of reviewers), review velocity (the rate and timing of reviews over the product's history) and review content quality (quality and authenticity of review text). In summary statistics in Table 1, we find that most of the products in data source 1 have grade A (Excellent) or grade B (Very Good).

The fourth data source is Keepa.[21] We download from Keepa the sales ranks of the products in data source 1. Keepa provides us with the sales rank history of around 70% of the products in data source 1. Following the classical paper in online reviews Chevalier and Mayzlin (2006), we use the logarithmic transformation of sales ranks as the dependent variable in our analysis. The history of featured offers is collected to check whether Amazon products collected from our first source (i.e., Amazon) are mostly sold by Amazon in the history and to run robustness checks.

Table 1: Summary Statistics

	Count	Mean	STD	25%	50%	75%
Data Source 1: Amazon						
Number of Products	16,921					
Number of Reviews	16,921	6,296.77	21,831.15	124	844	4074
Average Ratings	16,522	4.431	0.333	4.3	4.5	4.6
Price	16,921	175.42	810.05	20.99	42.50	139.99
Amazon-Sold Product	16,882	29.74%				
Data Source 2: Dataset Compiled from Multiple Research by Brett Hollenbeck						
Number of Products	15,041					
Fake Review Purchasing Products	13,146					
Number of Reviews	15,041	30.37	98.16	4	6	12
Average Ratings	15,041	4.290	0.870	3.941	4.667	5
Percentage of Fake Reviews	15,041	67.71%	35.39%	50%	80%	100%
Data Source 3: RateBud						
Number of Products	16,713					
Products with Grade A (Excellent)	9,985					
Products with Grade B (Very Good)	4,840					
Products with Grade C (Good)	1,319					
Products with Grade D (Fair)	276					
Products with Grade E (Poor)	0					
Products with Grade F (Caution)	293					
Data Source 4: Keepa						
Number of Products	11,450					
Average Sales Rank	11,450	62,927.69	282,918.92	2,997.24	12,736.27	45,090.67
Median Sales Rank	11,450	52,096.19	297,861.34	1,575.25	7,481.75	30,258.50

4.2 Proxy for Products with Fake Reviews

This paper faces an empirical challenge common to all existing research in fake reviews: we do not know for sure which review is fake. We address this issue in three ways.

1. **Amazon-sold products:** Amazon very rarely participates in review manipulation (He, Hollenbeck, and Proserpio, 2022), so we use Amazon-sold products as a proxy for products less likely to have fake reviews.

2. **Suspicious products predicted by RateBud:** We collect suspicious products from RateBud,[7] a commercial website that grades the credibility of reviews for products on Amazon. RateBud provides letter grades for the credibility of the reviews of products, with A meaning excellent, B meaning very good, C meaning good, D meaning fair, E meaning poor, and F meaning caution. A and B are largely free from fraud, while D, E and F are suspicious manipulators. C, in the middle, is somewhat ambiguous. In the main text, we compare products in C, D, E, and F with the other products. We report the results when we compare products in D, E, and F with the other products in Appendices A.3, B.3 and C.2 for our empirical evidence on textual features and overrepresentation (testing Hypotheses 1a and 1b), bias in sentiments of AI summary (testing Hypotheses 2a and 2b) and market distortions (testing Hypothesis 3), respectively. The results are largely consistent.

3. **Review manipulation in the literature:** We use the review manipulators in He et al. (2022) and fake reviewers in Feldman, Tosyali, and Overgoor (2025) as another proxy. An important issue with this practice is that those are fraudsters publicly identified since three years ago. Therefore, according to a joint analysis with the dataset of Hou et al. (2024), by 2023 around 36% of fake reviews had already been removed by Amazon. More should have been removed by now, so this may not be a good proxy if we analyze the data we recently collected from Amazon (data source 1 in Section 4.1), but it is ideal for the dataset compiled by Brett Hollenbeck (data source 2 in Section 4.1).

5 Empirical Results

In this section, we test the hypotheses that we have derived in Section 3.3. We first study the textual features of fake reviews to figure out whether they overlap with features favored by AI summary. We then test whether, more reviews of products with more fake reviews are represented in the sentiments of AI summary. Next, we study the summary paragraphs and the sentiment corresponding to each keyword to test whether review manipulators receive more positive AI summary. Finally, we check whether, after the introduction of AI summary, review manipulators experienced higher sales.

5.1 Textual Features And Overrepresentation of Fake Reviews

We first test Hypotheses 1a and 1b. We first study the textual features of fake reviews to see whether fake reviews are more similar to each other, more positive, and longer, compared with authentic reviews. The textual features imply that fake reviews are overrepresented in sentiments of AI summary. We therefore test whether reviews of fake review products are more represented in AI summary.

5.1.1 Textual Features of Fake Reviews

We analyze the dataset compiled by Brett Hollenbeck from different research papers. They download reviews from Amazon.[25] We classify a review as a fake review if it is (i) a 5-star review, (ii) written for a review manipulating seller identified by He et al. (2022), and (iii) written by a fake reviewer identified by Feldman, Tosyali, and Overgoor (2025). We compare the classified fake reviews with the other reviews.

We first test Hypothesis 1a by focusing on analyzing the review length, sentiment and textual similarities. To compare the length, we simply count the number of characters and words. To compare the sentiments of review texts, we use OpenAI GPT-4.1,[27] and set the temperature (i.e., the randomness in generating the results) to zero, so that the results are replicable. To compare similarities, we first convert sentences in reviews into embeddings using the popular model all-MiniLM-L6-v2. We also try other embedding models as a robustness check in Appendix A.1. We calculate the cosine similarities of the embeddings of reviews within the same product. We select

products with enough reviews (more than 40 reviews) and enough fake reviews (more than 20 fake reviews). We also try other thresholds as a robustness check in Appendix A.2.

The results are shown in Panel A of Table 2. The p-values come from Welch's t-test. We find that fake reviews are significantly more similar to each other, express significantly more positive sentiments, and are significantly longer with more words and more characters. All the results are significant no matter whether we compare with all authentic reviews or authentic 5-star reviews. The findings present clear evidence to support Hypothesis 1a.

To test Hypothesis 1b, we run keyword-extraction algorithms on the dataset to see whether a greater portion of fake reviews mention common themes. We run HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), which is a key part of open-source keyword-extraction algorithms such as KeyBERT and BERTopic.[28] In these algorithms, HDBSCAN is first applied to cluster texts that have common themes. Then, outliers that do not belong to any clusters are dropped, and other methods are applied to assign a keyword to each cluster. In our practice, we first select products with enough reviews (more than 40 reviews) and enough fake reviews (more than 20 fake reviews). Then, we run HDBSCAN to cluster reviews. Finally, we calculate what percentage of fake reviews and authentic reviews are in the clusters (i.e., mention the keywords). Robustness checks using other embeddings and thresholds are presented in Appendices A.1 and A.2, respectively. Panel B of Table 2 presents the results of the analysis of archived review texts. The results show that a significantly larger percentage of fake reviews mention the common keywords no matter whether we compare with all authentic reviews or authentic 5-star reviews. This supports our Hypothesis 1b.

5.1.2 Overrepresentation of Fake Reviews

One concern with the analysis above is that the methods we adopt to extract keywords may be different from what is actually used by Amazon. Therefore, we directly analyze the keywords extracted by Amazon with their own keyword-extraction algorithms. In particular, we study whether Amazon's algorithms find similar patterns in Panel B of Table 2.

Here, the AI summary keywords are downloaded directly from Amazon (data source 1 in Section 4.1). We use the first two proxies introduced in Section 4.2 to capture products with more

Table 2: Analysis of Review Texts with Archived Reviews

	Fake Reviews	Authentic Reviews	Authentic 5-Star Reviews
Panel A: Textual Features in Hypothesis 1a			
Length (# of Characters)	322.70	237.48 ($p < 0.001$)	212.55 ($p < 0.001$)
Length (# of Words)	61.28	44.62 ($p < 0.001$)	39.95 ($p < 0.001$)
Sentiment	4.66		4.57 ($p < 0.001$)
Cosine Similarity	0.2026	0.1743 ($p < 0.001$)	0.1889 ($p < 0.001$)
Panel B: Clustering in Hypothesis 1b			
% that Mention Keywords	50.35%	41.74% ($p < 0.001$)	44.09% ($p = 0.001$)

Note. This table compares fake versus all authentic and authentic 5-star Amazon reviews using an archived review dataset. The reported p-values are all from Welch's t-test with fake reviews. Panel A: Length counts characters and words in the review text. Sentiment is produced by OpenAI GPT-4.1. Cosine Similarity is the average within-product pairwise cosine similarity of embeddings (all-MiniLM-L6-v2). Panel B: Texts are clustered within product using HDBSCAN (as in common keyword-extraction pipelines). We compare the percentage of reviews that fall into the clusters (i.e., mention the extracted keywords).

or fewer fake reviews, i.e., products sold by Amazon and products with low grades from RateBud. Specifically, we regard C, D, E and F as low grades. In Appendix A.3, we present results with D, E and F as low grades. Amazon displays the number of reviews by which each keyword is mentioned. We regard this as the cluster size. We calculate average cluster size and minimum cluster size across keywords, and normalize both of them with the number of reviews. "Coverage Ratio" is what percentage of reviews are in clusters, so that they are input into sentiment analysis and thus represented in the sentiments of keywords and summary paragraphs.[29] It equals "# of Keywords $\times$ Average Cluster Size / # of Reviews".

The results are presented in Table 3. The p-values are from Welch's t-test. We compare Amazon-sold products and Non-Amazon products in Panel A, and fake review products predicted by RateBud and other products in Panel B. In Panel C, we conduct a stricter comparison. We compare fake review products predicted by RateBud and the other products within non-Amazon products. For products more likely to have fake reviews, all these measures consistently show that more reviews mention common keywords, i.e., the cluster size is larger. As a consequence, a greater portion of reviews contribute to the sentiment analysis and are therefore represented in AI summary.

One concern with this analysis is that there may be an upper bound of the number of keywords that can be displayed on the Amazon user interface. Products with fewer fake reviews may be more popular, so they have a greater number of reviews, but their number of keywords is bounded. Therefore, some of their common themes are dropped, which can explain why lower portion of their reviews mention common themes. In Appendix A.4, we report the results on the subset of

Table 3: Analysis of Keywords Extracted with Amazon's Algorithms

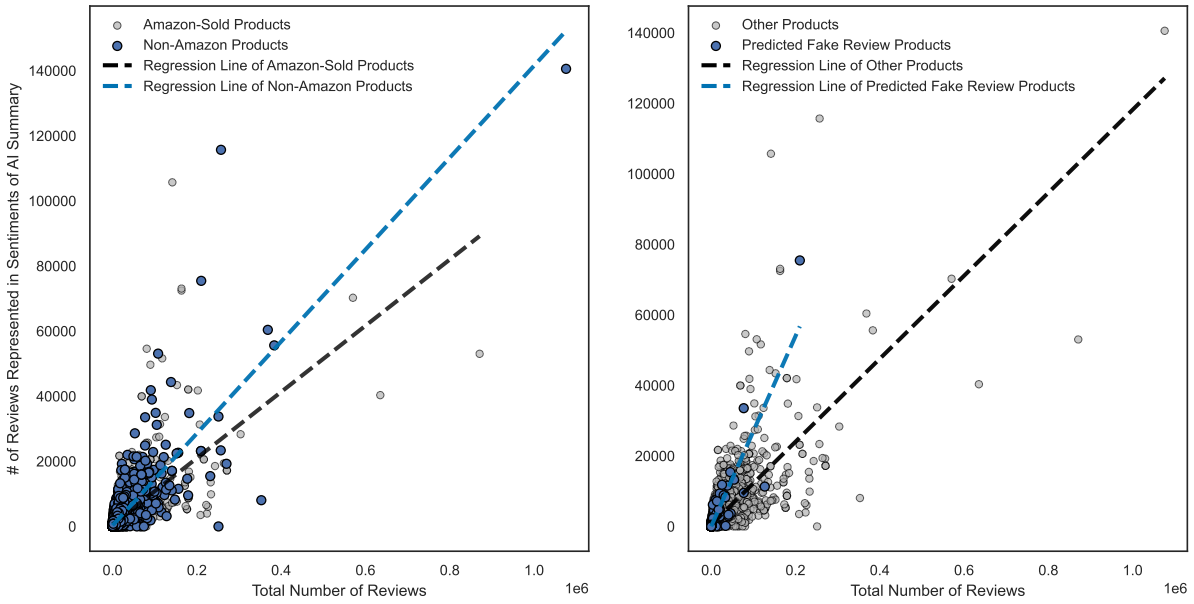
Panel A: Comparison between Amazon-Sold Products and Non-Amazon Products			
	Non-Amazon Products	Amazon-Sold Products	P-Value
Average Cluster Size/# of Reviews	0.0642	0.0454	<0.001
Minimum Cluster Size/# of Reviews	0.0338	0.0229	<0.001
Coverage Ratio	0.4528	0.3302	<0.001
Panel B: Comparison between Predicted Fake Review Products by RateBud and Other Products			
	Fake Review Products	Other Products	P-Value
Average Cluster Size/# of Reviews	0.1222	0.0551	<0.001
Minimum Cluster Size/# of Reviews	0.0774	0.0280	<0.001
Coverage Ratio	0.7423	0.3979	<0.001
Panel C: Comparison Predicted Fake Review Products and Other Non-Amazon Products			
	Fake Review Products	Other Non-Amazon Products	P-Value
Average Cluster Size/# of Reviews	0.1224	0.0608	<0.001
Minimum Cluster Size/# of Reviews	0.0770	0.0312	<0.001
Coverage Ratio	0.7495	0.4361	<0.001

Note: This table analyzes how Amazon's keyword extraction algorithms work on products with more vs fewer fake reviews. Cluster size is by how many reviews each keyword is mentioned. "Average Cluster Size" and "Minimum Cluster Size" are both aggregations of cluster sizes across keywords. "Coverage Ratio" is what percentage of reviews are in clusters so that they are represented in sentiments of keywords and summary paragraphs. It is "# of Keywords $\times$ Average Cluster Size / # of Reviews".

products that are not bounded by the upper bound. The results are largely consistent with what is shown in Table 3. Interestingly, after we switch to the subsample of products whose number of keywords are not bounded, the coverage ratio decreases more for products with more fake reviews, suggesting that the reviews of many fake review products are so clustered around some common themes.

Another concern is that the results of the comparison of the size of the cluster and the number of keywords in Table 3 are driven by the fact that we normalize them by dividing the number of reviews. To mitigate this concern, in Figure 2, we plot the number of reviews taken to sentiment analysis not normalized by the number of reviews (Coverage Ratio $\times$ # of Reviews). They are the number of reviews that are input into sentiment analysis, and are ultimately represented in the sentiments of keywords and AI summary paragraphs. It is clear in Figure 2 that the majority of products with more fake reviews are above the regression line of products with fewer fake reviews. Additionally, the regression line of products with more fake reviews is steeper than and is above the regression line of products with fewer fake reviews. This result shows that no matter what the total number of reviews is, AI summary systematically takes more reviews into sentiment analysis from the reviews of fake review products.

Figure 2: Overrepresentation of Fake Reviews



Note: This figure plots the number of reviews input into sentiment analysis and thus represented in the sentiments of AI summary, against the total number of reviews. The figure shows that for products with any total number of reviews, AI summary algorithms take more reviews of fake review products into sentiment analysis and represent them in the sentiments of AI summary.

A further concern is that Amazon may only take a fixed number of reviews to extract keywords from. However, it is evident from Figure 2 that the number of reviews that Amazon uses to extract keywords and analyze sentiments are constantly increasing in the total number of reviews without reaching an upper bar. Instead, the results are driven by the fact that for products with any total number of reviews, Amazon picks more reviews to summarize on products with more fake reviews.

One may also question whether the finding is because fake review sellers offer different categories of products. To mitigate this concern, we control for categories and search terms in Appendix A.5. The results are consistent with the findings in Table 3.

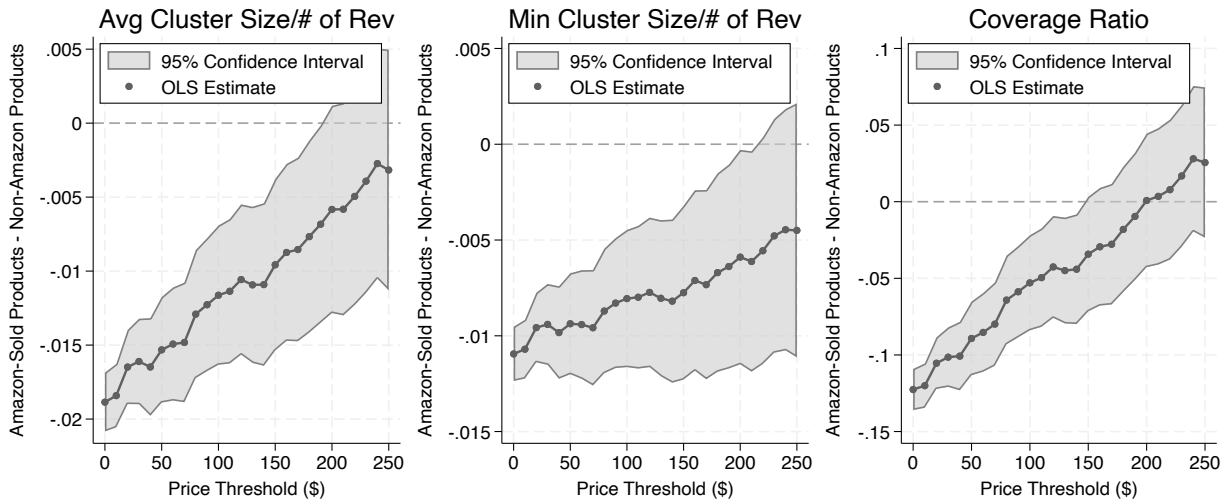
5.1.3 Falsification Test

There are a lot of possible differences between Amazon-sold and non-Amazon products. To further ensure that the difference of interests in fake reviews plays an important role in the findings, we leverage the nature of the market for fake reviews to conduct a falsification test. Amazon is a platform with relatively stricter regulation on review manipulation. Therefore, sellers who want to buy fake reviews typically have to cover the costs of fake reviewers to buy the product to ensure

that they are verified customers (He, Hollenbeck, and Proserpio, 2022). This implies that the cost of buying fake reviews increases with the price of the products. Therefore, we should expect that more expensive products should be less likely to have fake reviews. If our theory is correct, without fake reviews, AI summary should give no systematic bias between Amazon and non-Amazon products.[30]

We plot the difference (Amazon-Sold Products - Non-Amazon Products) in average cluster size/# of reviews, minimum cluster size/# of reviews, and coverage ratio in Panel A of Table 3 on the subsample of products with prices higher than a threshold. The difference in coverage ratio is roughly the difference between the slope of the blue line and the dark line in Figure 2. The results are reported in Figure 3. It is clear from the figure that the difference in the proportion of reviews represented in AI summary exists only when the products are fairly cheap, suggesting that the difference in fake reviews between Amazon-sold and non-Amazon products plays a critical role in driving the results.

Figure 3: Falsification Test for Overrepresentation



We conjecture that the difference in the number of reviews that are input into sentiment analysis (i.e., the gap between the blue line and the dark line in Figure 2) is dominated by fake reviews. If so, we should expect that sentiments of AI summary of review manipulators are more positive.

5.2 Bias in Sentiments of AI Summary

We test Hypotheses 2a and 2b here. Keyword extraction and AI summary are designed to search for common themes across texts, so they favor fake reviews that mention more common themes and fewer idiosyncratic personal details. Additionally, as we have shown earlier, fake reviews are more positive than authentic reviews even after controlling for the number of stars in the rating. As a result, we should expect that products with more fake reviews have: (i) more positive sentiments for extracted keywords, (ii) more positive AI summary paragraphs, and (iii) larger differences between the sentiments (measured on the scale of ratings) of summary paragraphs and average ratings.

5.2.1 Model Specification

The general idea of the empirical analysis is straightforward. We want to compare three dependent variables between products more (or less) likely to have fake reviews and other products: sentiments associated with keywords, sentiments (measured on the scale of ratings) of summary paragraphs, and the difference between sentiments (measured on the scale of ratings) of summary paragraphs and average ratings. The sentiment associated with each keyword is directly observable on Amazon (the checkmark and minus signs before each keyword shown in Figure 1). We assign 1 to positive sentiment (checkmark sign), -1 to negative sentiment (minus sign), and 0 to neutral sentiment (no sign). Then, we take the average across different keywords. To convert summary paragraphs to sentiments measured on the scale of ratings, we use two language models: OpenAI GPT-4.1[27] and bert-base-multilingual-uncased-sentiment.[31] When using OpenAI GPT-4.1, we prompt the model to predict quality perceived by consumers who see the summary paragraph, on the scale of average ratings. bert-base-multilingual-uncased-sentiment is a specialized language model finetuned on the task of predicting the ratings of reviews, so the output is designed to fall within the scale of ratings. We report the results with OpenAI GPT-4.1 in the main text, and the results with bert-base-multilingual-uncased-sentiment in Appendix B.1.

The analysis faces an empirical challenge. There may be other differences between subgroups with more (or less) fake reviews that drive the results, especially between Amazon-sold products and non-Amazon products, because Amazon decides which product to sell by itself. We use three

methods to deal with this concern. First, we flexibly control the distribution of ratings. Second, we control the characteristics of products by fixed effects, including the categories of the products and search term × page number fixed effects. In this way, we only compare among products on the same search result page. Third, we conduct a falsification test in Section 5.2.3. We find that the difference between Amazon-sold products and non-Amazon products is concentrated in cheaper products for which review manipulation is prevalent and disappears for expensive products whose reviews are more costly to manipulate.

The specification we estimate is shown in Equation (1),

$$DV_i = \beta I_i + Rating\_Controls_i\gamma + \phi_c + \psi_{kn} + \epsilon_i, \tag{1}$$

where i denotes products, c denotes categories, k denotes search terms, and n denotes the page number in the search results on which product i is shown. DV_i is one of the three dependent variables that we just mentioned, including sentiments corresponding to keywords, sentiments (measured on the scale of ratings) of summary paragraphs, and the difference between sentiments (measured on the scale of ratings) of summary paragraphs and actual average ratings, I_i is the indicator for products with more (or less) fake reviews based on the three proxies discussed in Section 4.2, $Rating\_Controls_i$ is a vector of flexible controls of ratings. ϕ_c is category fixed effects and ψ_{kn} is search term × page number fixed effects. They are included to control for unobservable characteristics of the products and to ensure that our comparison is within the products shown on the same search result page.

5.2.2 Results

We first consider I_i to indicate whether product i is sold by Amazon. The results are shown in Table 4. The results of the sentiments corresponding to keywords are reported in Columns (1)-(3), the results of the sentiments (measured on the scale of ratings) of the summary paragraphs are reported in Columns (4)-(6), and the results of the difference between the sentiments (measured on the scale of ratings) of the summary paragraphs and the actual average ratings are reported in Columns (7)-(9). We find that with search term × page number fixed effects and category fixed effects controlled, Amazon-sold products (unlikely to have fake reviews) have significantly more

negative sentiments associated with keywords, significantly more negative sentiments of summary paragraphs, and significantly smaller difference between sentiment (measured on the scale of ratings) of AI summary and average ratings than the other products on the same search result page. R^2 of the regressions is considerable, suggesting that we have successfully controlled for many factors that can influence the dependent variables. Results in Columns (1)-(6) validate Hypothesis 2a, while results in Columns (7)-(9) validate Hypothesis 2b.

Table 4: Sentiment Comparison between Amazon-Sold Products and Other Products

	(1) Keyword	(2) Keyword	(3) Keyword	(4) Paragraph	(5) Paragraph	(6) Paragraph	(7) Difference	(8) Difference	(9) Difference
Amazon-Sold	-0.0447***	-0.0443***	-0.0443***	-0.0751***	-0.0744***	-0.0739***	-0.0743***	-0.0748***	-0.0738***
	(0.00479)	(0.00477)	(0.00477)	(0.00868)	(0.00866)	(0.00866)	(0.00871)	(0.00871)	(0.00866)
Avg Rating	0.693***	0.820***	0.839***	1.150***	1.347***	1.066***			
	(0.00848)	(0.0167)	(0.0393)	(0.0154)	(0.0303)	(0.0713)			
# of Reviews	-3.79e-07***	-3.54e-07***	-3.57e-07***	-9.59e-07***	-9.21e-07***	-8.89e-07***	-8.66e-07***	-9.16e-07***	-8.83e-07***
	(9.16e-08)	(9.14e-08)	(9.15e-08)	(1.66e-07)	(1.66e-07)	(1.66e-07)	(1.67e-07)	(1.67e-07)	(1.66e-07)
Variance of Ratings		0.0619***	0.0643***		0.0954***	0.0572***		-0.0296***	0.0455***
		(0.00699)	(0.00850)		(0.0127)	(0.0154)		(0.00645)	(0.00889)
Share of 5-Star Reviews			-0.0397			0.617***			0.735***
			(0.0782)			(0.142)			(0.0602)
Constant	-2.367***	-3.020***	-3.075***	-0.902***	-1.909***	-1.070***	-0.238***	-0.196***	-0.850***
	(0.0376)	(0.0827)	(0.135)	(0.0682)	(0.150)	(0.244)	(0.00420)	(0.0101)	(0.0545)
Search Term × Page No. FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	13569	13569	13569	13569	13569	13569	13569	13569	13569
R^2	0.553	0.556	0.556	0.532	0.534	0.535	0.295	0.296	0.304

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We then focus the analysis on non-Amazon products which are more likely to have fake reviews. We use RateBud to identify a subgroup of products that are even more likely to have fake reviews. This is the stricter comparison in Panel C of Table 3. We group C, D, E, and F together as fake review products in Table 5. In Appendix B.3, we report the results with only D, E and F as fake review products as a robustness check. In both of them, all three dependent variables are significantly larger for fake review products. This also validates our hypotheses.

Next, we put the two group indicators, predicted fake review products and Amazon-sold products, together in one specification. Table 6 reports the results, which are consistent with those presented in Tables 4 and 5.

5.2.3 Falsification Test

There are a variety of differences between Amazon-sold products and non-Amazon products. To further ensure that the difference in fake reviews rather than other differences plays an important

Table 5: Sentiment Comparison between Predicted Fake Review Products and Other Non-Amazon Products

	(1) Keyword	(2) Keyword	(3) Keyword	(4) Paragraph	(5) Paragraph	(6) Paragraph	(7) Difference	(8) Difference	(9) Difference
Fake	0.0371***	0.0399***	0.0422***	0.0941***	0.0990***	0.0954***	0.0893***	0.0898***	0.0911***
	(0.0102)	(0.0102)	(0.0102)	(0.0185)	(0.0184)	(0.0185)	(0.0185)	(0.0185)	(0.0184)
Avg Rating	0.662***	0.802***	0.903***	1.110***	1.353***	1.191***			
	(0.0105)	(0.0212)	(0.0503)	(0.0191)	(0.0383)	(0.0908)			
# of Reviews	-5.75e-07***	-5.52e-07***	-5.65e-07***	-1.20e-06***	-1.16e-06***	-1.13e-06***	-1.12e-06***	-1.14e-06***	-1.11e-06***
	(1.31e-07)	(1.31e-07)	(1.31e-07)	(2.37e-07)	(2.37e-07)	(2.37e-07)	(2.37e-07)	(2.38e-07)	(2.37e-07)
Variance of Ratings		0.0669***	0.0812***		0.116***	0.0932***		-0.0109	0.0588***
		(0.00880)	(0.0109)		(0.0159)	(0.0197)		(0.00792)	(0.0109)
Share of 5-Star Reviews			-0.214*			0.344*			0.676***
			(0.0966)			(0.175)			(0.0736)
Constant	-2.230***	-2.945***	-3.253***	-0.724***	-1.966***	-1.471***	-0.240***	-0.224***	-0.825***
	(0.0466)	(0.105)	(0.174)	(0.0843)	(0.190)	(0.315)	(0.00413)	(0.0122)	(0.0666)
Search Term × Page No. FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9214	9214	9214	9214	9214	9214	9214	9214	9214
R^2	0.565	0.568	0.568	0.548	0.551	0.552	0.333	0.333	0.340

Standard errors in parentheses

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

role in the findings, we again leverage the nature of the market for fake reviews to conduct a falsification test. The cost of buying fake reviews increases with the price of the products. Therefore, we should expect that more expensive products should be less likely to have fake reviews. If our theory is correct, without fake reviews, AI summary should give no systematic bias between Amazon and non-Amazon products.[30]

We rerun the specifications in Columns (3), (6) and (9) of Table 4,[32] on the subsample with prices higher than a threshold, and keep track of the coefficients of the Amazon-sold products group indicator. The results are shown in Figure 4. As is evident, the negative coefficients are significantly negative only when the prices of the products are fairly small. In the subsample with products more expensive than $100 and thus fake reviews are very unlikely, the effects we have discovered in Table 4 no longer exist, suggesting that fake reviews are likely to be the reason underlying the difference in the three dependent variables between Amazon-sold products and non-Amazon products.

5.3 AI Summary Benefits Review Manipulators

5.3.1 Specification

We test Hypothesis 3. More specifically, we investigate whether, after the introduction of AI summary, sales of products with more fake reviews increased. Following the practice of the classical paper in the literature of online reviews, Chevalier and Mayzlin (2006), we use the logarithms of

Table 6: Specification with Both Group Indicators

	(1) Keyword	(2) Keyword	(3) Keyword	(4) Paragraph	(5) Paragraph	(6) Paragraph	(7) Difference	(8) Difference	(9) Difference
Fake	0.0435***	0.0469***	0.0479***	0.106***	0.112***	0.106***	0.0973***	0.0994***	0.102***
	(0.00893)	(0.00890)	(0.00895)	(0.0161)	(0.0161)	(0.0162)	(0.0162)	(0.0162)	(0.0161)
Amazon-Sold	-0.0442***	-0.0435***	-0.0436***	-0.0750***	-0.0740***	-0.0737***	-0.0742***	-0.0748***	-0.0735***
	(0.00481)	(0.00480)	(0.00480)	(0.00871)	(0.00869)	(0.00868)	(0.00875)	(0.00874)	(0.00868)
Avg Rating	0.697***	0.836***	0.876***	1.160***	1.378***	1.149***			
	(0.00854)	(0.0169)	(0.0399)	(0.0154)	(0.0307)	(0.0722)			
# of Reviews	-3.73e-07***	-3.46e-07***	-3.51e-07***	-9.43e-07***	-9.01e-07***	-8.75e-07***	-8.45e-07***	-8.97e-07***	-8.62e-07***
	(9.17e-08)	(9.15e-08)	(9.15e-08)	(1.66e-07)	(1.66e-07)	(1.66e-07)	(1.66e-07)	(1.67e-07)	(1.66e-07)
Variance of Ratings		0.0666***	0.0721***		0.105***	0.0739***		-0.0311***	0.0474***
		(0.00706)	(0.00862)		(0.0128)	(0.0156)		(0.00646)	(0.00893)
Share of 5-Star Reviews			-0.0876			0.501***			0.768***
			(0.0789)			(0.143)			(0.0607)
Constant	-2.390***	-3.097***	-3.217***	-0.949***	-2.066***	-1.380***	-0.241***	-0.197***	-0.880***
	(0.0379)	(0.0839)	(0.137)	(0.0685)	(0.152)	(0.248)	(0.00430)	(0.0102)	(0.0549)
Search Term × Page No. FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	13416	13416	13416	13416	13416	13416	13416	13416	13416
R^2	0.555	0.559	0.559	0.536	0.538	0.539	0.298	0.299	0.308

Standard errors in parentheses

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

sales ranks as our dependent variable. We collect sales rank data from 2022 to 2024 from Keepa. We select products that have observations both before and after the introduction of AI summary. We aggregate the observations of ranks to the daily level to make the panel more balanced.

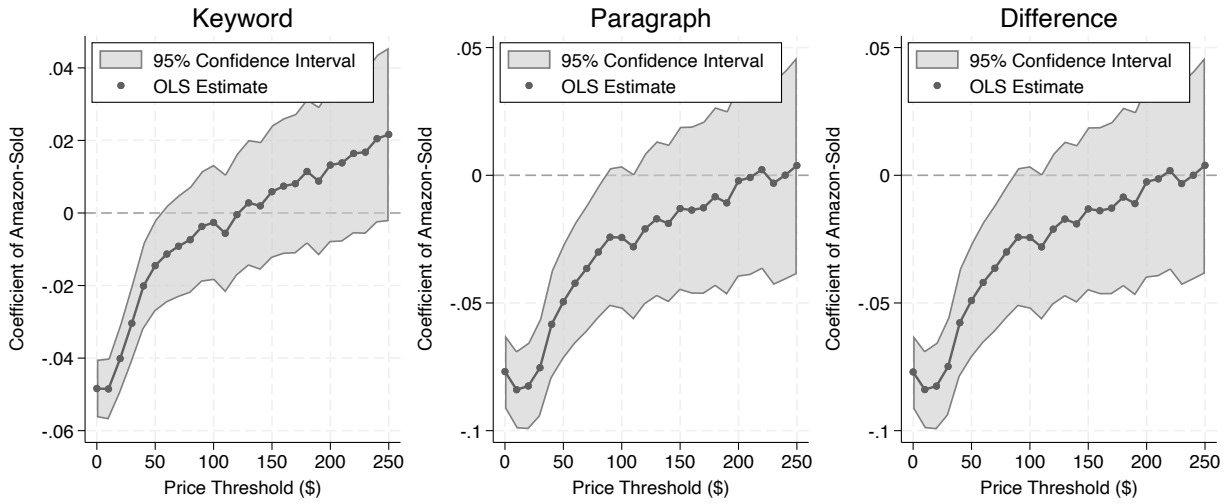
We estimate a model with the specification in Equation (2), where i denotes products, and t denotes date. I_i is again the indicators for groups of products that are more (or less) likely to have fake reviews. To make the groups of products comparable, we run exact matching on categories and the interaction of search terms and page numbers. Following the suggestions by Bertrand, Duflo, and Mullainathan (2004), we cluster the error term at the product level to correct for the serial autocorrelation across time.

$$\log(Sales\_Ranks_{it}) = \beta I_i \times After_t + \alpha_i + \omega_t + \epsilon_{it} \tag{2}$$

We also conduct event studies in Equation (3). Specifically, we replace $After_t$ with a dummy variable for each month ϕ_m. We also run exact matching on categories and the interaction of search terms and page numbers. We cluster the error term at the product level.

$$\log(Sales\_Ranks_{it}) = \sum_m \beta_m I_i \times \phi_m + \alpha_i + \omega_t + \epsilon_{it} \tag{3}$$

Figure 4: Falsification Test for Bias in Sentiments



5.3.2 Results

The results are reported in Table 7. In columns (1) and (2), we compare the change in sales between Amazon-sold products and non-Amazon products. Regardless of whether we match the data or not, compared with non-Amazon products, Amazon-sold products, which are less likely to have fake reviews, suffered from an approximately 10% increase in sales ranks after the introduction of AI summary. The results of the robustness check with Amazon share are reported in Appendix C.1.

In columns (3) and (4), we report the comparison between predicted fake review products and other products. We group C, D, E, and F together as fake review products in the main analysis. In the Appendix C.2, we report the results with only D, E and F as fake review products as a robustness check. No matter whether we match or not, the sales ranks of predicted fake reviews products moved up by around 20% after the introduction of AI summary. We notice that the coefficients for predicted fake review products are larger and more significant after matching.

We report the results of event studies. We plot the coefficients for the interactions of each month dummy and group indicator. Following common practice, we normalize the month when AI summary was introduced as Month 0 and normalize the coefficients of Month -1 to be 0. We report the results with matching in the main text, and the results without matching in Appendix C.3. The result of the comparison between Amazon-sold and other products is shown in Figure 5, while the result of the comparison between predicted fake review and other products is shown

Table 7: Comparison in Sales Change after the Introduction of AI Summary

	(1) Log(Rank)	(2) Log(Rank)	(3) Log(Rank)	(4) Log(Rank)
Amazon-Sold × After	0.113***	0.103*		
	(0.0293)	(0.0476)		
Fake × After			-0.186*	-0.227**
			(0.0770)	(0.0863)
Constant	8.544***	8.488***	8.569***	9.115***
	(0.00564)	(0.0106)	(0.00115)	(0.00386)
Product FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Matched on Category	No	Yes	No	Yes
Matched on Search Term × Page No.	No	Yes	No	Yes
Observations	6336696	4812712	6336696	2099141
R^2	0.805	0.814	0.805	0.809

Standard errors in parentheses

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

in Figure 6. There were no significant pretrends before the introduction of AI summary. Instead, there was a significant break in the evolution of β_m after the introduction of AI summary. The sales ranks of products with more fake reviews tend to relatively move up, suggesting that their sales relatively increased following the introduction of AI summary. In Appendix C.3, we find consistent results even if we do not match the subgroups on categories and keywords. There is no significant pretrend but a significant break in the evolution of β_m.

In conclusion, we observe that following the introduction of AI Summary, the sales of products changed in a direction that supports market distortion stated in Hypothesis 3.

6 Conclusion

Fake reviews have been shown to harm consumer welfare (He, Hollenbeck, and Proserpio, 2022; Gandhi, Hollenbeck, and Li, 2024), and have received much attention from the Federal Trade Commission (FTC) of the US[4] and the Parliament of the UK.[5] In this paper, we find evidence that the negative effects of fake reviews can be amplified by AI summary of reviews, which has been a common practice on digital platforms. This AI bias is fundamentally rooted in the design of AI algorithms. AI summary is designed to summarize common themes among user-generated reviews, while more fake reviews mention common themes. Sentiment analysis built in AI summary can

Figure 5: Event Study for the Comparison between Amazon-Sold and Other Products

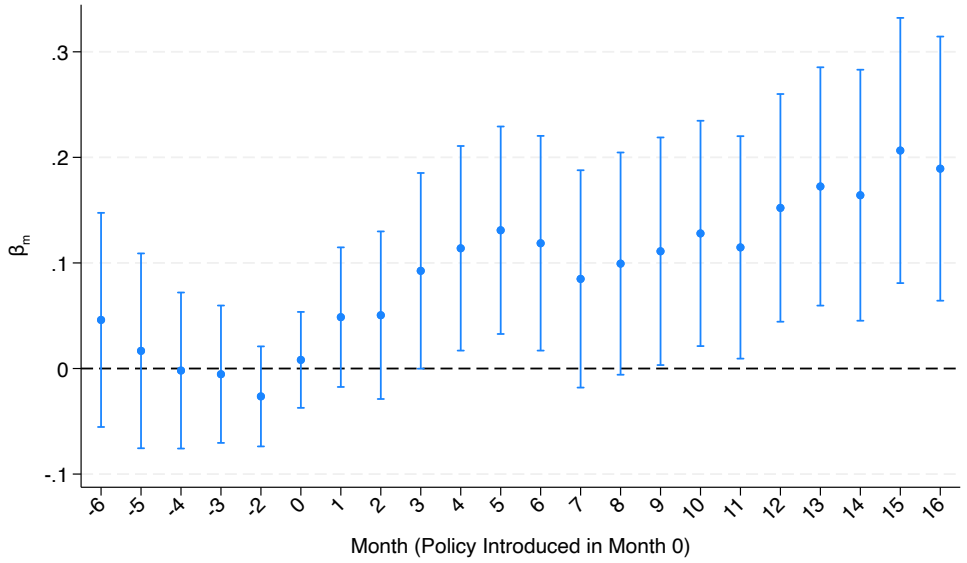
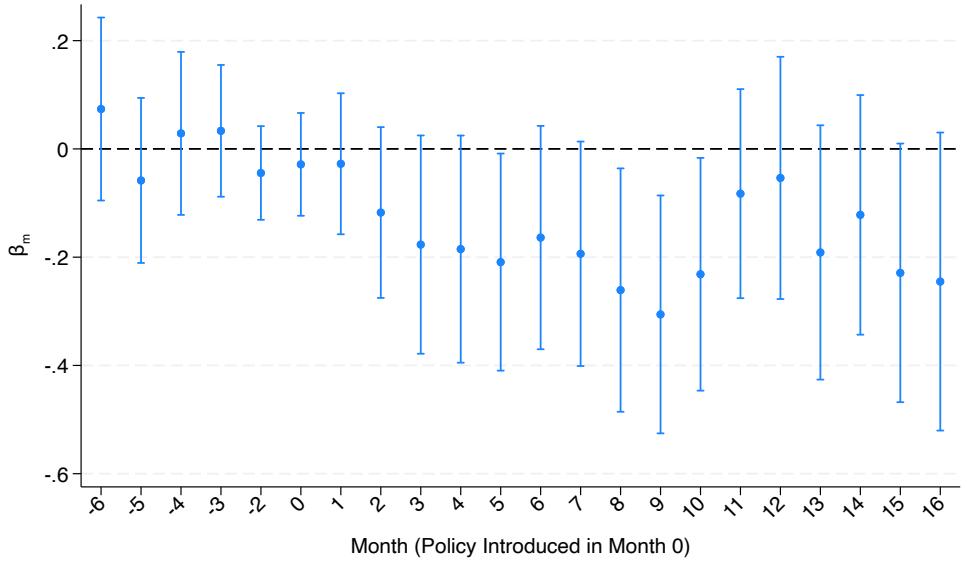


Figure 6: Event Study for the Comparison between Predicted Fake and Other Products



extract sentiments beyond ratings from review texts, while fake reviews are more positive than authentic reviews, and even authentic 5-star reviews. The more positive sentiments in fake reviews can thus be readily reflected in AI summary.

We first show that fake reviews are longer, more positive, and more similar to each other. We run keyword-extraction algorithms ourselves to demonstrate that more fake reviews mention common themes. We also analyze the keywords extracted by Amazon. We find that for fake review products, more reviews are assigned to keywords and thus reflected in the sentiments of AI summary. Then, we show that the sentiments associated with keywords and AI summary paragraphs are more positive for products that are more likely to have fake reviews, even after we control for flexible metrics of average ratings and abundant fixed effects. Finally, we study the economic effects on the market. The introduction of AI summary benefits review manipulators more than the other sellers.

This distortive effect on the market is particularly concerning to society. Researchers have shown that AI summary can increase purchase rates (Wang, Tong, and Dong, 2025) and hence benefit platforms. This explains why platforms are increasingly adopting AI summary. However, our findings suggest that AI summary might direct consumers to suboptimal products due to the interaction between the mechanics of their algorithm and the nature of fake reviews. It is plausible that this might hurt consumers. Investigating its impact on consumer welfare is an important research question that we leave for future research.

In addition, the findings are worrying for platforms that care about the long-term trust of customers and also have interests in deploying AI tools. We find that current AI summary might overrepresent fake reviews. In the long run, this might hurt customer trust in the review system. Therefore, a responsible platform that truly cares about long-term customer satisfaction should be alert to the unexpected AI bias of its algorithms. When deploying trendy AI tools on their platforms, managers should be aware that AI is new and complicated, so it can lead to unexpected and unintended outcomes that drive customers away.

It is challenging to design AI summary algorithms that can summarize honest common themes while avoiding common themes in false information. Simply improving the accuracy of the current algorithms only makes the bias even worse. Nevertheless, we hope this research will inspire more computer scientists and researchers to work towards algorithms proficient in summarizing information contaminated by misinformation, such as online reviews.

Notes

[1] https://www.wsj.com/tech/personal-tech/fake-reviews-and-inflated-ratings-are-still-a-problem-for-amazon-11623587313
[2] https://apnews.com/article/fake-online-reviews-generative-ai-40f5000346b1894a778434ba295a0496
[3] https://www.wsj.com/us-news/youre-probably-falling-for-fake-product-reviews-b4d07f23
[4] https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials
[5] https://www.legislation.gov.uk/ukpga/2024/13/contents
[6] https://www.aboutamazon.com/news/amazon-ai/amazon-improves-customer-reviews-with-generative-ai
[7] https://www.ratebud.ai/
[8] https://apnews.com/article/amazon-generative-ai-reviews-products-4db85ac68c0d46f7b728b9da48da1a96
[9] https://corporate.walmart.com/news/2025/06/06/walmart-the-future-of-shopping-is-agentic-meet-sparky
[10] https://www.nngroup.com/articles/ai-reviews
[11] https://techcrunch.com/2024/12/10/yelp-adds-ai-powered-review-insights-to-restaurants/
[12] https://www.theverge.com/news/624891/ai-generated-review-summaries-coming-to-apples-app-store
[13] https://www.androidauthority.com/play-store-ai-generated-review-summaries-3611995/
[14] Collected from https://www.amazon.com/dp/B0DWJCK9LZ/ on November 1, 2025
[15] Collected from https://www.bestbuy.com/product/apple-11-inch-ipad-air-m3-chip-built-for-apple-intelligence-wi-fi-128gb-blue/JJGCQ8VZVZ on November 1, 2025
[16] https://www.emarketer.com/content/5-charts-search-2024-google-ai-retail-media
[17] https://aws.amazon.com/blogs/machine-learning/going-beyond-ai-assistants-examples-from-amazon-com-reinventing-industries-with-generative-ai/
[18] In our downloaded data, the total number of mentions of keywords can exceed the total number of reviews.
[19] https://www.channelmax.net/article/amazon-introduces-ai-generated-review-summaries-to-help-customers-shop-faster-and-smarter
[20] https://nymag.com/intelligencer/2022/07/amazon-fake-reviews-can-they-be-stopped.html
[21] https://keepa.com/
[22] https://explodingtopics.com/blog/most-searched-items-on-amazon
[23] https://meetglimpse.com/top-searched/most-searched-products-on-amazon/
[24] We also stop collecting data when we have already collected more than 110 products but not reached page seven.
[25] https://github.com/bretthollenbeck/fake-reviews-data
[26] We collected the number of reviews of those products on September 29, 2025. The current average number of reviews is 24,019.28
[27] https://openai.com/index/gpt-4-1/
[28] https://www.nlplanet.org/course-practical-nlp/02-practical-nlp-first-tasks/12-clustering-articles
[29] If one review aligns with two or more keywords, we also count it multiple times, because it contributes to the sentiments corresponding to multiple keywords, and is thus more represented than a review that only aligns with one keyword.
[30] Similar exercise is not applicable to comparison of sales, because even if AI summary of expensive products is not influenced at all, their sales will be impacted by cheaper competitors whose AI summary is influenced.
[31] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment
[32] We omit fixed effects in subsample analysis.

References

Acemoglu, D., Como, G., Fagnani, F., and Ozdaglar, A. (2013). Opinion fluctuations and disagreement in social networks. *Mathematics of Operations Research*, 38(1):1–27.

Acemoglu, D., Ozdaglar, A., and ParandehGheibi, A. (2010). Spread of (mis)information in social networks. *Games and Economic Behavior*, 70(2):194–227.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

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

Cowgill, B. and Tucker, C. E. (2019). Economics, fairness and algorithmic bias. *preparation for: Journal of Economic Perspectives*.

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science*, 52(10):1577–1593.

Farronato, C. and Zervas, G. (2022). Consumer reviews and regulation: Evidence from nyc restaurants. *National Bureau of Economic Research*.

Feldman, E., Tosyali, A., and Overgoor, G. (2025). Addressing large-scale reviewer recruitment on amazon: A reviewer-centric approach to the fake review problem. *SSRN working paper No. 5156231*.

Gandhi, A., Hollenbeck, B., and Li, Z. (2024). Misinformation and mistrust: The equilibrium effects of fake reviews on amazon.com. *Working Paper*.

González-Bailón, S., Lazer, D., Barberá, P., Zhang, M., Allcott, H., Brown, T., Crespo-Tenorio, A., Freelon, D., Gentzkow, M., Guess, A. M., et al. (2023). Asymmetric ideological segregation in exposure to political news on facebook. *Science*, 381(6656):392–398.

He, S., Hollenbeck, B., Overgoor, G., Proserpio, D., and Tosyali, A. (2022). Detecting fake-review buyers using network structure: Direct evidence from amazon. *Proceedings of the National Academy of Sciences*, 119(47):e2211932119.

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

Hou, Y., Li, J., He, Z., Yan, A., Chen, X., and McAuley, J. (2024). Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining - WSDM '08*, page 219. ACM Press.

Lambrecht, A. and Tucker, C. (2019). Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65(7):2966–2981.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–870.

Li, J., Ott, M., Cardie, C., and Hovy, E. (2014). Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Liu, Y. and Lapata, M. (2019). Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Lu, T., Yuan, M., Wang, C., and Zhang, X. M. (2022). Histogram distortion bias in consumer choices. *Management Science*, 68:8963–8978.

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427.

Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., and Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences*, 118(23).

Mayzlin, D. (2006). Promotional chat on the internet. *Marketing Science*, 25(2):155–163.

Mayzlin, D., Dover, Y., and Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8):2421–2455.

Mostagir, M., Ozdaglar, A., and Siderius, J. (2022). When is society susceptible to manipulation? *Management Science*, 68(10):7153–7175.

Mukherjee, A., Liu, B., and Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, pages 191–200.

Nyhan, B., Settle, J., Thorson, E., Wojcieszak, M., Barberá, P., Chen, A. Y., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., et al. (2023). Like-minded sources on facebook are prevalent but not polarizing. *Nature*, 620(7972):137–144.

Obermeyer, Z. and Mullainathan, S. (2019). Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 89–89.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., and Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592:590 – 595.

Rayana, S. and Akoglu, L. (2015). Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 985–994. ACM.

Robertson, R. E., Green, J., Ruck, D. J., Ognyanova, K., Wilson, C., and Lazer, D. (2023). Users choose to engage with more partisan news than they are exposed to on google search. *Nature*, 618(7964):342–348.

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., and Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, 9(1):4787.

Shapira, O., Pasunuru, R., Dagan, I., and Amsterdamer, Y. (2021). Multi-document keyphrase extraction: Dataset, baselines and review. *arXiv preprint arXiv:2110.01073*.

Su, Y., Wang, Q., Rhee, K., and Qiu, L. (2025). Less to process, more to express: The impact of AI-generated summaries on review diversity. *Working Paper*.

Sun, M. (2012). How does the variance of product ratings matter? *Management Science*, 58(4):696–707.

Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS ONE*, 13.

Vicario, M. D., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113:554 – 559.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wang, S., Tong, J., and Dong, J. Q. (2025). When generative artificial intelligence meets human reviews: Effects on consumer behavior and hotel sales. *Working Paper*.

Wu, C., Che, H., Chan, T. Y., and Lu, X. (2015). The economic value of online reviews. *Marketing Science*, 34(5):739–754.

Zhu, F. and Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing*, 74(2):133–148.

Zhu, X. and Pechmann, C. C. (2024). Political polarization triggers conservatives' misinformation spread to attain ingroup dominance. *Journal of Marketing*, 89(1):39–55.

Appendix A Robustness Checks for Textual Features of Fake Reviews

A.1 Other Embedding Models

We rerun the analyses with a new embedding model named multi-qa-MiniLM-L6-cos-v1. As the name suggests, this model is more specialized in the analysis of cosine similarity, and in capturing the similarities among texts. The results are shown in Table S.1. The relative relationship is exactly the same as the analysis in the main text, although review texts are easier to be clustered with the more specialized embedding.

A.2 Other Thresholds

We try a stricter threshold here. Specifically, we select products with more than 60 reviews and more than 30 fake reviews to ensure that we have enough reviews to analyze for each product. The results are shown in Table S.2. The relative relationship remains consistent with what is presented in the main text.

A.3 Other Cutoffs

RateBud provides letter grades for the credibility of reviews of products, with A meaning excellent, B meaning very good, C meaning good, D meaning fair, E meaning poor, and F meaning caution. In the main text, we report the results with C, D, E and F as predicted fake review products. Here, we rerun the analyses with D, E and F as predicted fake review products. However, products with grades D, E and F are scarce in our data, because we only collected products shown on the first several pages of the search results. The results are shown in Table S.3, which are largely consistent with the results in Table 3.

A.4 Results on the Subset Not Bounded

A concern with the analysis in Table 3 is that there exists an upper bound of the number of keywords that can be displayed on the user interface on Amazon. In our dataset, the maximum number of keywords displayed is 8. Therefore, we select the subset of products whose number of keywords does not reach the upper bound (the number of keywords smaller than 8) and rerun the analyses.

The results are presented in Table S.4. The results are not only consistent with the results in Table 3, but also statistically very significant. Interestingly, after we switch to the subsample of products whose number of keywords are not bounded, the coverage ratio decreases more for products with more fake reviews, suggesting that the reviews of many fake review products are so clustered around some common themes.

A.5 With Controls for Categories

One may think the results shown in Table 3 are because fake review sellers are selling different products than other sellers. Here, we control for the categories and search terms of products using regression analysis. The results are shown in Table S.5.

Appendix B Robustness Checks for Bias in Sentiments of AI Summary

B.1 Sentiment Analysis with Specialized BERT

In the main text, we use the famous general-purpose model OpenAI GPT-4.1 to convert summary paragraphs to sentiment measured on the scale of ratings. Here, we use a specialized BERT model finetuned to predict ratings of reviews, named bert-base-multilingual-uncased-sentiment,[31] to convert summary paragraphs to sentiments measured on the scale of ratings. The output of the model is designed to fall on a 1–5 scale. The results are shown in Tables S.6, S.7 and S.8.

B.2 Share of Time in Buy Box

A tricky point here is that products sold by Amazon when we collected the data may not always be sold by Amazon. On Amazon, different sellers of a product share the same product page and compete to be the featured offer prominently displayed in the Buy Box.[33] Amazon pools different sellers of one product together. To deal with this concern, we collect the history of the featured offer in Buy Box from Keepa. We confirm that during 80% of the time, the Buy Boxes of Amazon-sold products in our data were occupied by Amazon, while during only 6% of the time, the Buy Boxes of the other products were occupied by Amazon. This history is only available for a subset of products, so we still use Amazon-sold products as the proxy in the main text.

We present the results with the share of time during which the Buy Box of each product is occupied by Amazon as a proxy for how unlikely the products have fake reviews. We replace Amazon-sold indicator with the continuous Amazon share variable. The result is presented in Table S.9. The result is largely consistent with that shown in the main text. Products less likely to have fake reviews (with larger Amazon share) have significantly more negative sentiment corresponding to keywords, summary paragraphs and difference between sentiment and average ratings.

B.3 Other Cutoffs

RateBud provides letter grades for the credibility of reviews of products, with A meaning excellent, B meaning very good, C meaning good, D meaning fair, E meaning poor, and F meaning caution. In the main text, we report the results with C, D, E and F as predicted fake review products. Here, we rerun the analyses with D, E and F as predicted fake review products. However, products with grades D, E and F are scarce in our data, because we only collected products shown on the first several pages of the search results.

We compare fake review products with a stricter cutoff (C is not fake) and other non-Amazon products. The results are presented in Table S.10. Results are still significant and consistent with what is shown in the main text, in spite of the smaller size of fake review products.

B.4 Products with Fake Reviews Predicted by He et al. (2022)

In this section, we compare the products predicted by He et al. (2022) and the products we have collected. However, this comparison has three tricky issues. First, according to a joint analysis with the dataset of Hou et al. (2024), by 2023 around 36% of fake reviews of those products had been removed, so the remaining fake reviews now may not be large. Second, these products existed before 2022 and were found to engage in review manipulation, so those who survive even now must be the good ones among them. Third, these products have at least 3 years of history, so they have significantly more reviews. We need to use flexible controls of the number of reviews.

We collect data from Amazon for products predicted by He et al. (2022) to have fake reviews, and then we compare those products with products we collected using the popular keywords. The results are shown in Table S.11. They also support Hypotheses 2a and 2b.

Appendix C Robustness Checks for AI Summary Benefits Review Manipulators

C.1 Share of Time in Buy Box

A tricky point here is that products sold by Amazon when we collected the data may not always be sold by Amazon. On Amazon, different sellers of a product share the same product page and compete to be the featured offer prominently displayed in the Buy Box.[34] Amazon pools different sellers of one product together. To deal with this concern, we collect the history of the featured offer in Buy Box from Keepa. We confirm that during 80% of the time, the Buy Boxes of Amazon-sold products in our data were occupied by Amazon, while during only 6% of the time, the Buy Boxes of the other products were occupied by Amazon. This history is only available for a subset of products, so we still use Amazon-sold products as the proxy in the main text.

We replace the indicator for Amazon-sold products with the share of time during which the Buy Box is occupied by Amazon as a proxy for how unlikely the products have fake reviews. We then rerun the analysis in the main text. The results are presented in Columns (1) and (2) of Table S.12, which are largely consistent with our results in the main text. Products less likely to have fake reviews (with a greater Amazon share) experienced a decline in sales after the introduction of AI summary.

C.2 Other Cutoffs

RateBud provides letter grades for the credibility of reviews of products, with A meaning excellent, B meaning very good, C meaning good, D meaning fair, E meaning poor, and F meaning caution. In the main text, we report the results with C, D, E and F as predicted fake review products. Here, we rerun the analysis with D, E and F as predicted fake review products. The results are presented in Columns (3) and (4) of Table S.12. The results are still largely consistent with those in the main text.

C.3 Event Study without Matching

Here, we report the results of event studies without matching. The result of the comparison between Amazon-sold and other products is presented in Figure S.1, while the result of the comparison between predicted fake review and other products is presented in Figure S.2. Interestingly, we find no significant pretrends even if we do not match the subgroups on categories and keywords.

Appendix D Appendix Figures and Tables

Table S.1: Analysis of Review Texts with multi-qa-MiniLM-L6-cos-v1

	Fake Reviews	Authentic Reviews	Authentic 5-Star Reviews
Panel A: Textual Features in Hypothesis 1a			
Cosine Similarity	0.2105	0.1913 ($p < 0.001$)	0.2070 ($p < 0.001$)
Panel B: Clustering in Hypothesis 1b			
% that Mention Keywords	66.92%	62.86% ($p < 0.001$)	65.36% ($p = 0.292$)

Table S.2: Analysis of Review Texts with a Stricter Threshold

	Fake Reviews	Authentic Reviews	Authentic 5-Star Reviews
Panel A: Textual Features in Hypothesis 1a			
Cosine Similarity	0.2024	0.1749 ($p < 0.001$)	0.1985 ($p < 0.001$)
Panel B: Clustering in Hypothesis 1b			
% that Mention Keywords	49.65%	41.09% ($p < 0.001$)	43.08% ($p = 0.002$)

Table S.3: Analysis of Review Texts with Amazon's Algorithms with C as Authentic

Panel B: Comparison between Predicted Fake Review Products and Other Products			
	Fake Review Products	Other Products	P-Value
Average Cluster Size/# of Reviews	0.1586	0.0578	<0.001
Minimum Cluster Size/# of Reviews	0.1084	0.0299	<0.001
Coverage Ratio	0.8973	0.4117	<0.001
Panel C: Comparison within Non-Amazon Products between Predicted Fake Review Products and Other Products			
	Fake Review Products	Other Non-Amazon Products	P-Value
Average Cluster Size/# of Reviews	0.1521	0.0638	<0.001
Minimum Cluster Size/# of Reviews	0.1031	0.0334	<0.001
Coverage Ratio	0.8868	0.4509	<0.001

Table S.4: Analysis of Review Texts with Amazon's Algorithms on the Unbounded Subsample

Panel A: Comparison between Amazon-Sold Products and Non-Amazon Products			
	Non-Amazon Products	Amazon-Sold Products	P-Value
Average Cluster Size/# of Reviews	0.0820	0.0681	<0.001
Minimum Cluster Size/# of Reviews	0.0533	0.0435	<0.001
Coverage Ratio	0.3506	0.2931	<0.001
Panel B: Comparison between Predicted Fake Review Products and Other Products			
	Fake Review Products	Other Products	P-Value
Average Cluster Size/# of Reviews	0.1399	0.0709	<0.001
Minimum Cluster Size/# of Reviews	0.1029	0.0444	<0.001
Coverage Ratio	0.5481	0.3099	<0.001
Panel C: Comparison within Non-Amazon Products between Predicted Fake Review Products and Other Products			
	Fake Review Products	Other Non-Amazon Products	P-Value
Average Cluster Size/# of Reviews	0.1363	0.0744	<0.001
Minimum Cluster Size/# of Reviews	0.0997	0.0469	<0.001
Coverage Ratio	0.5386	0.3247	<0.001

Table S.5: Analysis of Keywords Extracted by Amazon with Controls for Categories

	(1) Average	(2) Average	(3) Min	(4) Min	(5) Coverage Ratio	(6) Coverage Ratio
Fake	0.0646***	0.0471***	0.0480***	0.0371***	0.327***	0.216***
	(0.00220)	(0.00206)	(0.00156)	(0.00153)	(0.0150)	(0.0136)
Amazon-Sold	-0.0169***	-0.0148***	-0.00951***	-0.00800***	-0.114***	-0.103***
	(0.000990)	(0.00111)	(0.000702)	(0.000827)	(0.00678)	(0.00736)
Constant	0.0607***	0.0616***	0.0311***	0.0315***	0.435***	0.444***
	(0.000571)	(0.000534)	(0.000405)	(0.000398)	(0.00391)	(0.00354)
Search Term × Page No. FE	No	Yes	No	Yes	No	Yes
Category FE	No	Yes	No	Yes	No	Yes
Observations	14306	13379	14306	13379	14306	13379
R^2	0.079	0.440	0.077	0.381	0.053	0.462

Standard errors in parentheses

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

Figure S.1: Event Study for the Comparison between Amazon-Sold and Other Products without Matching

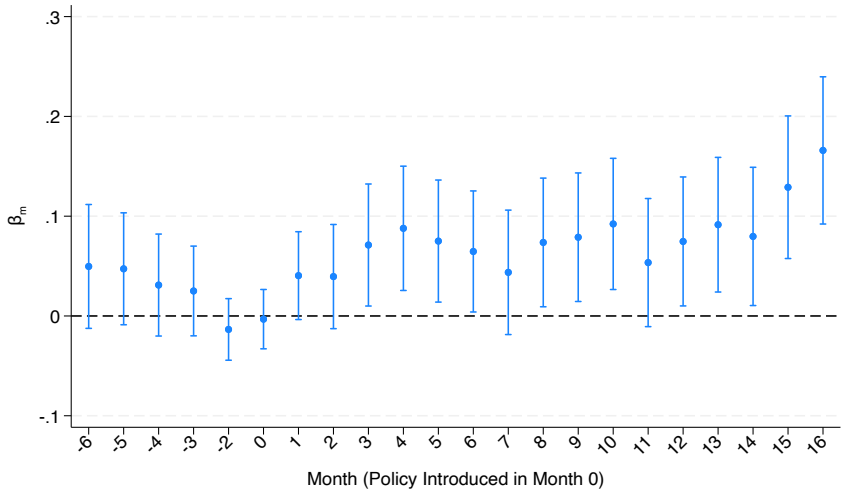


Table S.6: Sentiment Comparison between Amazon-Sold Products and Non-Amazon Products with BERT

	(1) Paragraph	(2) Paragraph	(3) Paragraph	(4) Difference	(5) Difference	(6) Difference
Amazon-Sold	-0.0817***	-0.0815***	-0.0806***	-0.0806***	-0.0819***	-0.0810***
	(0.0114)	(0.0114)	(0.0114)	(0.0115)	(0.0114)	(0.0114)
Avg Rating	1.198***	1.247***	0.732***			
	(0.0202)	(0.0399)	(0.0939)			
# of Reviews	-1.27e-06***	-1.26e-06***	-1.20e-06***	-1.15e-06***	-1.26e-06***	-1.23e-06***
	(2.18e-07)	(2.18e-07)	(2.18e-07)	(2.19e-07)	(2.19e-07)	(2.18e-07)
Variance of Ratings		0.0237	-0.0464*		-0.0654***	0.000867
		(0.0167)	(0.0203)		(0.00846)	(0.0117)
Share of 5-Star Reviews			1.131***			0.649***
			(0.187)			(0.0793)
Constant	-1.737***	-1.988***	-0.449	-0.859***	-0.766***	-1.343***
	(0.0896)	(0.198)	(0.322)	(0.00552)	(0.0133)	(0.0717)
Search Term × Page No. FE	Yes	Yes	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	13569	13569	13569	13569	13569	13569
R^2	0.434	0.434	0.436	0.237	0.240	0.245

Standard errors in parentheses

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

Table S.7: Sentiment Comparison between Predicted Fake Review Products and Other Non-Amazon Products with BERT

	(1) Paragraph	(2) Paragraph	(3) Paragraph	(4) Difference	(5) Difference	(6) Difference
Fake	0.157***	0.159***	0.151***	0.148***	0.151***	0.152***
	(0.0244)	(0.0244)	(0.0245)	(0.0245)	(0.0245)	(0.0244)
Avg Rating	1.191***	1.288***	0.918***			
	(0.0252)	(0.0509)	(0.121)			
# of Reviews	-1.78e-06***	-1.77e-06***	-1.72e-06***	-1.65e-06***	-1.75e-06***	-1.73e-06***
	(3.14e-07)	(3.14e-07)	(3.14e-07)	(3.15e-07)	(3.15e-07)	(3.14e-07)
Variance of Ratings		0.0466*	-0.00588		-0.0572***	0.00899
		(0.0211)	(0.0262)		(0.0105)	(0.0145)
Share of 5-Star Reviews			0.786***			0.642***
			(0.232)			(0.0977)
Constant	-1.715***	-2.213***	-1.082**	-0.873***	-0.790***	-1.361***
	(0.112)	(0.252)	(0.418)	(0.00547)	(0.0161)	(0.0884)
Search Term × Page No. FE	Yes	Yes	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9214	9214	9214	9214	9214	9214
R^2	0.456	0.456	0.457	0.277	0.279	0.283

Standard errors in parentheses

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

Table S.8: Specification with Both Group Indicators and BERT

	(1) Paragraph	(2) Paragraph	(3) Paragraph	(4) Difference	(5) Difference	(6) Difference
Fake	0.167***	0.169***	0.158***	0.155***	0.160***	0.162***
	(0.0212)	(0.0212)	(0.0213)	(0.0213)	(0.0212)	(0.0212)
Amazon-Sold	-0.0799***	-0.0795***	-0.0789***	-0.0789***	-0.0802***	-0.0790***
	(0.0114)	(0.0114)	(0.0114)	(0.0115)	(0.0115)	(0.0114)
Avg Rating	1.210***	1.285***	0.842***			
	(0.0203)	(0.0404)	(0.0951)			
# of Reviews	-1.25e-06***	-1.23e-06***	-1.19e-06***	-1.12e-06***	-1.23e-06***	-1.20e-06***
	(2.18e-07)	(2.18e-07)	(2.18e-07)	(2.19e-07)	(2.19e-07)	(2.18e-07)
Variance of Ratings		0.0362*	-0.0245		-0.0667***	0.00348
		(0.0169)	(0.0206)		(0.00848)	(0.0118)
Share of 5-Star Reviews			0.969***			0.686***
			(0.188)			(0.0800)
Constant	-1.797***	-2.181***	-0.852**	-0.866***	-0.771***	-1.381***
	(0.0901)	(0.200)	(0.326)	(0.00565)	(0.0134)	(0.0724)
Search Term × Page No. FE	Yes	Yes	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	13416	13416	13416	13416	13416	13416
R^2	0.438	0.439	0.440	0.241	0.245	0.249

Standard errors in parentheses

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

Table S.9: Robustness Check with Share of Time in Buy Box

	(1) Keyword	(2) Keyword	(3) Keyword	(4) Paragraph	(5) Paragraph	(6) Paragraph	(7) Difference	(8) Difference	(9) Difference
Amazon Share	-0.0702***	-0.0694***	-0.0694***	-0.114***	-0.112***	-0.112***	-0.111***	-0.112***	-0.112***
	(0.00581)	(0.00579)	(0.00579)	(0.0105)	(0.0105)	(0.0105)	(0.0106)	(0.0106)	(0.0105)
Avg Rating	0.701***	0.829***	0.806***	1.165***	1.364***	1.000***			
	(0.00880)	(0.0172)	(0.0408)	(0.0159)	(0.0312)	(0.0738)			
# of Reviews	-3.01e-07**	-2.75e-07**	-2.72e-07**	-8.33e-07***	-7.93e-07***	-7.52e-07***	-7.31e-07***	-7.89e-07***	-7.52e-07***
	(9.29e-08)	(9.26e-08)	(9.27e-08)	(1.68e-07)	(1.68e-07)	(1.68e-07)	(1.69e-07)	(1.69e-07)	(1.68e-07)
Variance of Ratings		0.0622***	0.0590***		0.0969***	0.0479**		-0.0342***	0.0480***
		(0.00721)	(0.00876)		(0.0131)	(0.0159)		(0.00669)	(0.00921)
Share of 5-Star Reviews			0.0521			0.810***			0.809***
			(0.0822)			(0.149)			(0.0628)
Constant	-2.402***	-3.057***	-2.988***	-0.964***	-1.984***	-0.903***	-0.236***	-0.186***	-0.905***
	(0.0390)	(0.0853)	(0.139)	(0.0706)	(0.155)	(0.252)	(0.00454)	(0.0107)	(0.0568)
Search Term × Page No. FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	12804	12804	12804	12804	12804	12804	12804	12804	12804
R^2	0.560	0.562	0.562	0.541	0.543	0.544	0.306	0.307	0.317

Standard errors in parentheses

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

Table S.10: Sentiment Comparison between Predicted Fake Review Products and Other Non-Amazon Products with C as Authentic

	(1) Keyword	(2) Keyword	(3) Keyword	(4) Paragraph	(5) Paragraph	(6) Paragraph	(7) Difference	(8) Difference	(9) Difference
Fake	0.0490^{+}	0.0528^{+}	0.0557^{+}	0.146**	0.152**	0.142**	0.141**	0.142**	0.141**
	(0.0295)	(0.0294)	(0.0295)	(0.0534)	(0.0533)	(0.0534)	(0.0535)	(0.0535)	(0.0533)
Avg Rating	0.652***	0.759***	0.829***	1.090***	1.277***	1.013***			
	(0.00948)	(0.0195)	(0.0455)	(0.0172)	(0.0353)	(0.0824)			
# of Reviews	-7.34e-07***	-7.09e-07***	-7.19e-07***	-1.48e-06***	-1.44e-06***	-1.40e-06***	-1.41e-06***	-1.44e-06***	-1.40e-06***
	(1.22e-07)	(1.22e-07)	(1.22e-07)	(2.21e-07)	(2.21e-07)	(2.21e-07)	(2.21e-07)	(2.22e-07)	(2.21e-07)
Variance of Ratings		0.0518***	0.0615***		0.0903***	0.0533**		-0.0123^{+}	0.0509***
		(0.00826)	(0.0101)		(0.0150)	(0.0182)		(0.00729)	(0.0103)
Share of 5-Star Reviews			-0.150^{+}			0.570***			0.593***
			(0.0889)			(0.161)			(0.0689)
Constant	-2.186***	-2.733***	-2.943***	-0.638***	-1.591***	-0.795**	-0.239***	-0.221***	-0.751***
	(0.0419)	(0.0967)	(0.157)	(0.0759)	(0.175)	(0.285)	(0.00401)	(0.0113)	(0.0627)
Search Term × Page No. FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9749	9749	9749	9749	9749	9749	9749	9749	9749
R^2	0.501	0.504	0.504	0.480	0.482	0.483	0.236	0.236	0.243

Standard errors in parentheses

$^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table S.11: Sentiment Comparison between Fake Review Products Predicted by He et al. (2022) and Other Products

	(1) Keyword	(2) Keyword	(3) Keyword	(4) Paragraph	(5) Paragraph	(6) Paragraph	(7) Difference	(8) Difference	(9) Difference
Fake (He et al., 2022)	0.0235**	0.0273***	0.0263***	0.0186	0.0322*	0.0315*	0.00733	0.0280*	0.0321*
	(0.00767)	(0.00771)	(0.00770)	(0.0141)	(0.0141)	(0.0141)	(0.0141)	(0.0141)	(0.0141)
Avg Rating	0.697***	0.610***	0.855***	1.165***	0.855***	1.050***			
	(0.00764)	(0.0210)	(0.0365)	(0.0140)	(0.0384)	(0.0669)			
# of Reviews	-1.10e-06***	-1.08e-06***	-1.09e-06***	-2.50e-06***	-2.43e-06***	-2.44e-06***	-2.17e-06***	-2.49e-06***	-2.43e-06***
	(1.35e-07)	(1.35e-07)	(1.34e-07)	(2.47e-07)	(2.47e-07)	(2.47e-07)	(2.47e-07)	(2.46e-07)	(2.46e-07)
# of Reviews^2	4.24e-12***	4.16e-12***	4.23e-12***	8.65e-12***	8.37e-12***	8.42e-12***	7.73e-12***	8.59e-12***	8.39e-12***
	(6.15e-13)	(6.15e-13)	(6.14e-13)	(1.13e-12)	(1.13e-12)	(1.13e-12)	(1.13e-12)	(1.13e-12)	(1.12e-12)
# of Reviews^3	-3.31e-18***	-3.25e-18***	-3.29e-18***	-6.12e-18***	-5.90e-18***	-5.93e-18***	-5.51e-18***	-6.06e-18***	-5.91e-18***
	(5.16e-19)	(5.16e-19)	(5.15e-19)	(9.46e-19)	(9.45e-19)	(9.44e-19)	(9.49e-19)	(9.44e-19)	(9.44e-19)
Share of 5-Star Reviews		0.261***	-0.0822		0.936***	0.664***		0.556***	0.753***
		(0.0592)	(0.0723)		(0.108)	(0.133)		(0.0394)	(0.0551)
Variance of Ratings			0.0641***			0.0508***			0.0421***
			(0.00780)			(0.0143)			(0.00822)
Constant	-2.396***	-2.208***	-3.129***	-0.990***	-0.317**	-1.046***	-0.261***	-0.675***	-0.881***
	(0.0339)	(0.0544)	(0.125)	(0.0622)	(0.0996)	(0.229)	(0.00392)	(0.0295)	(0.0499)
Category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	16366	16366	16366	16366	16366	16366	16366	16366	16366
R^2	0.527	0.528	0.530	0.505	0.507	0.508	0.257	0.267	0.268

Standard errors in parentheses

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

Table S.12: Robustness Check in Sales Analysis with Share of Time in Buy Box

	(1) Log(Rank)	(2) Log(Rank)	(3) Log(Rank)	(4) Log(Rank)
Amazon Share × After	0.550***	0.539***		
	(0.0314)	(0.0569)		
Fake × After			0.198	0.167
			(0.285)	(0.296)
Constant	8.466***	8.394***	8.566***	9.104***
	(0.00571)	(0.0124)	(0.000434)	(0.00141)
Product FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Matched on Category	No	Yes	No	Yes
Matched on Search Term × Page No.	No	Yes	No	Yes
Observations	6336696	4812712	6336696	2099141
R^2	0.807	0.816	0.805	0.809

Standard errors in parentheses

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

Figure S.2: Event Study for the Comparison between Predicted Fake Review and Other Products without Matching