AI Summaries Overrepresent Fake Reviews: Evidence from Amazon

Sihan Zhai[*] and Andrew T. Ching[†]

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

March 27, 2026

Abstract

AI summarization of reviews has been widely deployed to distill information from large volumes of consumer reviews. In this research, we argue that AI summarization produces an unintended consequence: AI summaries overrepresent fake reviews. Our key insights are that (i) AI summarization is designed to extract common themes, and that (ii) fake reviews are more linguistically similar to one another, creating more common themes. To provide empirical support for our argument, we study AI summaries on Amazon. Using varying assumptions and proxies for products more likely to contain fake reviews, we consistently find evidence that AI-generated summaries on Amazon overrepresent fake reviews. We also study the influence of the overrepresentation of fake reviews on market outcomes. We find that review manipulators receive more positive AI summaries than other sellers, even after we control average ratings and other observable characteristics. We also find that review manipulators experience a greater increase in sales following the introduction of AI summarization relative to their competitors.

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

1 Introduction

With the development of AI technologies, digital platforms (e.g., Amazon, Walmart, Best Buy, Yelp, Apple App Store, and Google Play Store) have recently launched AI summarization of reviews. According to Vaughn Schermerhorn, the director of community shopping at Amazon, AI summarization is intended to help consumers grasp common themes more easily from the vast volume of reviews on the platform.[1] AI summaries typically have two features: (i) a natural language processing model to extract common topics mentioned in the reviews, and (ii) sentiment analysis around each of the topics. A short paragraph is then composed based on the extracted topics and their average sentiments. Crucially, not all reviews are treated equally in this process. Specifically, reviews that mention more extracted topics contribute more to AI summaries, while reviews that do not mention common keywords extracted by summarization algorithms tend to be ignored.

Although this process of generating AI summaries seems reasonable, user reviews often consist of fake reviews (Mayzlin et al., 2014). Major media outlets—such as the Wall Street Journal and the Associated Press—repeatedly warn that fake reviews are widespread,[2,3] and that they can even fool experienced shoppers.[4] This concern has prompted regulatory and legislative action, including a recent rule issued by the Federal Trade Commission (FTC) in the US[5] and legislation enacted by the UK Parliament,[6] to combat fake reviews.

Building on this institutional background, we argue that AI summaries can unintentionally overrepresent fake reviews. This occurs because fake reviews tend to emphasize generic product merits and thus contain more common themes; simultaneously, AI summarization algorithms are explicitly designed to extract these prevalent patterns. Consequently, AI summaries may systematically overrepresent fraudulent content. Furthermore, because the tone of fake reviews remains more positive than authentic ones even after controlling for star ratings, this overrepresentation renders the summaries of manipulated products overly positive, potentially boosting sales of review manipulators.

In this paper, we define the overrepresentation of fake reviews as a condition where, on average, a fake review contains more extracted common themes than an authentic one. Building on this definition and our arguments above, we hypothesize that AI summaries overrepresent fake reviews,

leading to two distinct market effects: (i) review manipulators receive more positive AI summaries than other sellers, and (ii) review manipulators experience a greater increase in sales following the introduction of AI summaries relative to their competitors.

To test our hypotheses regarding the overrepresentation of fake reviews in AI summaries and its subsequent market effects, we assemble a dataset from multiple sources: Amazon, Keepa, and RateBud. We use Amazon-sold products as a proxy for items less likely to contain fake reviews, as Amazon rarely engages in fraudulent review manipulation (He et al., 2022b). Conversely, products with low review-credibility grades on RateBud[7] serve as a proxy for products with a higher prevalence of fake reviews. We further supplement our data by leveraging fake and authentic reviews identified in prior literature (He et al., 2022a; Feldman et al., 2025).

It is challenging to empirically test the overrepresentation of fake reviews because we cannot observe how many extracted keywords each review mentions. We adopt two methods to tackle this empirical challenge. First, we check the number of common themes each review mentions. Various algorithms consistently show that fake reviews mention more extracted common themes, which implies that AI summaries overrepresent fake reviews. Second, we conduct a product-level analysis based on AI summaries on Amazon. We find that products with a higher likelihood of containing fake reviews also tend to have a greater average number of mentions of the extracted common themes per review. This provides further support for our overrepresentation hypothesis.

Next, we test the predicted effects of overrepresentation on consumer information and market outcomes. First, we examine whether products with fake reviews receive systematically more positive AI summaries. Using OpenAI's GPT-4.1 and a fine-tuned BERT model specialized for review sentiment, we analyze both the extracted keywords (which correspond to common themes) and the short AI summary paragraph. We find evidence that products with a higher likelihood of containing fake reviews receive more positive AI summaries, both in terms of keyword sentiment and the overall content of the summary paragraph. This suggests that the overrepresentation of fake reviews fundamentally distorts market information. Furthermore, we test whether this algorithmic bias translates into distorted market outcomes. We find that following the introduction of AI summarization, the sales ranks of Amazon-sold products (which are significantly less susceptible to manipulation) dropped by approximately 10%. In contrast, products with low review-credibility grades on RateBud saw an average improvement of 20% in their sales ranks.

This paper contributes to three primary streams of literature. First, we document a new form of AI bias, arising as an unintended consequence of the interaction between the mechanistic nature of keyword extraction in AI summarization algorithms and the specific textual features of fake reviews. Second, we contribute to the literature on fake reviews by demonstrating that AI summaries can heighten the salience of fraudulent information, thereby improving the sales ranks of manipulated products, a shift that potentially harms consumer welfare. Finally, we add to the literature on misinformation transmission by showing that AI can function as an amplifier of misinformation, specifically in the context of deceptive fake product reviews.

The findings of this paper hold significant implications for policymakers. If platforms continue to deploy AI summarization of reviews without addressing the systematic bias documented here, the resulting market distortions could lead to substantial welfare loss for consumers. Furthermore, our research is critically relevant to platform managers focused on long-term customer trust; by inadvertently directing consumers toward sub-standard products, current AI summarization algorithms may fuel customer dissatisfaction and erode brand equity. Ultimately, we hope this work prompts AI developers to refine algorithms so they can more robustly summarize information in environments contaminated by misinformation.

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 institutional background. Section 4 develops a theoretical framework that defines the overrepresentation of fake reviews and proposes hypotheses. Section 5 introduces data sources. Section 6 reports the empirical evidence that AI summaries overrepresent fake reviews. Section 7 presents empirical results that review manipulators receive more positive AI summaries and that the introduction of AI summarization increases sales of review manipulators. Section 8 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. The literature on AI bias and algorithmic bias has attracted attention from scholars in various fields. Cowgill and Tucker (2019) provide a comprehensive literature review of algorithmic bias. In this paper, they distinguish algorithmic bias caused by biased objectives from algorithmic

bias caused by biased predictions, highlighting the particular challenges in dealing with the former. Solving algorithmic bias caused by biased objectives is a governance problem, which involves leadership, ethics, and oversight, rather than pure technical tweaks. A classic example is the filter bubble. Researchers (Levy, 2021; Nyhan et al., 2023; González-Bailón et al., 2023) have confirmed that recommendation algorithms expose users to more like-minded content that is consistent with users' own political beliefs. Other examples include Obermeyer and Mullainathan (2019), in which they find racial disparities in a commercial health risk prediction algorithm that is mistakenly optimized for healthcare cost rather than actual health. For another example, Lambrecht and Tucker (2019) find that STEM job information is more likely to be allocated to a male audience by recommendation algorithms. Our research contributes by documenting a new example of AI bias induced by biased objectives. We show that AI summarization favors fake reviews because it is unintentionally designed to seek textual features typically possessed by fake reviews.

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 et al., 2025) and sellers (Farronato and Zervas, 2022), and consequently influence consumer welfare (Wu et al., 2015). Therefore, fake reviews have raised widespread concerns. Mayzlin et al. (2014) are among the first to show convincing evidence of the existence of fake reviews, and Luca and Zervas (2016) investigate the economic incentives underlying review manipulation. He et al. (2022b) contribute significantly to the literature by directly observing the market for fake reviews, and therefore more confidently identifying fake reviews. Not only do they provide details of the market for fake reviews, they also shed light on whether review manipulation hurts consumer welfare (Dellarocas, 2006; Mayzlin, 2006). Gandhi et al. (2024) further study the welfare implications of fake reviews. Our paper contributes to this stream of literature by suggesting that the negative welfare impact of fake reviews might be reinforced by AI summaries in the AI era.

Our paper is related to the literature on the transmission of misinformation. Vosoughi et al. (2018) provide strong empirical evidence that misinformation diffuses significantly faster, farther, deeper, and more than authentic information. Researchers are interested in the reasons. In the theoretical literature, Acemoglu et al. (2010), Acemoglu et al. (2013) and Mostagir et al. (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 et al., 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 the 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 Institutional Background

3.1 AI Summary of Reviews

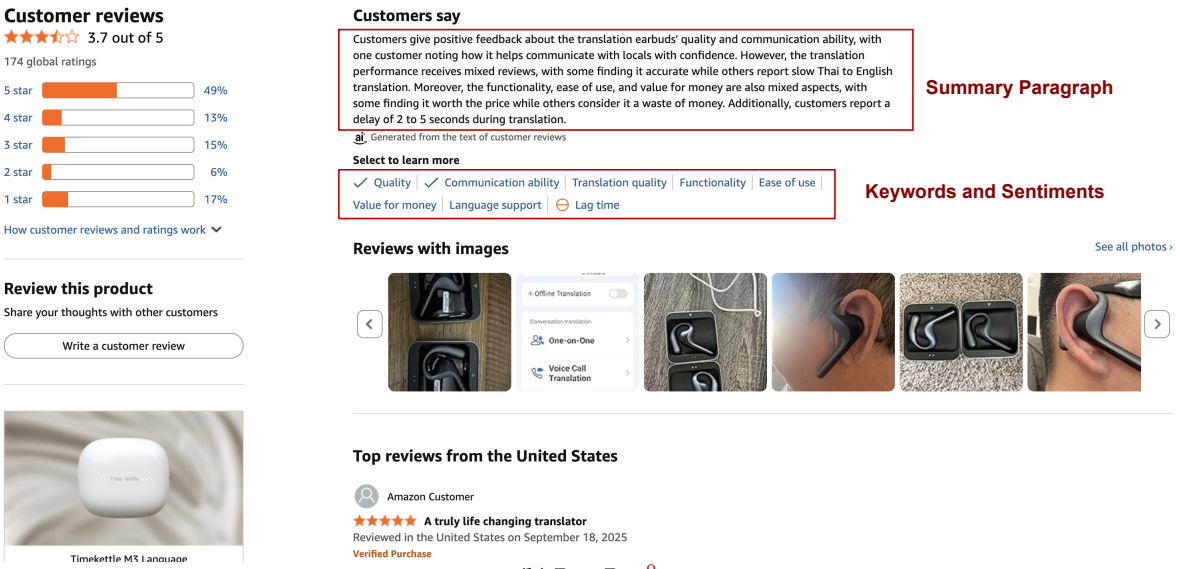
With the development of AI technologies, most digital platforms, including Amazon, Walmart, Best Buy, Yelp, Apple App Store and Google Play Store, have adopted AI summarization of reviews. AI summaries of reviews are machine-produced and language-model–based condensations of user-generated reviews that aim to mitigate information overload and support consumer decision-making on digital platforms.

AI summaries are generally displayed at a prominent location above all user-generated reviews. For example, Figure 1 shows how AI summaries are displayed on Amazon and Best Buy. AI summaries typically include 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 others 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.

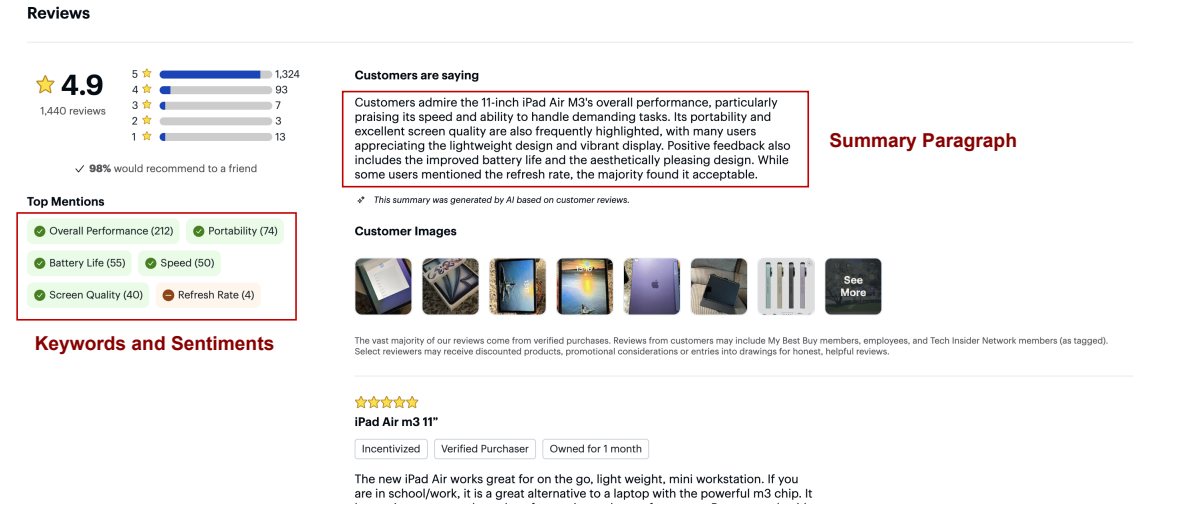
In this paper, we focus on Amazon. More than half (56%) of consumers in the US start their shopping searches on Amazon.[10] Amazon introduced AI summarization of reviews on August 14, 2023.[1] Regarding the business goal of AI summarization, Vaughn Schermerhorn, the director of community shopping at Amazon, claimed that they "want to make it easier for customers to understand the common themes across reviews." Moreover, he mentioned that "(AI summarization) provides a short paragraph right on the product detail page that highlights the product features

Figure 1: User Interfaces of AI Summaries

(a) Amazon[8]



(b) Best Buy[9]



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 the internal AI team of Amazon describes the basic features of the algorithm.[11] According to Amazon engineers, the AI summarization algorithm "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 that are most commonly mentioned across reviews. Amazon only uses reviews that mention this set of the most popular keywords for sentiment analysis. The sentiment analysis evaluates the tone of each review that

mentions the extracted keyword (by coding it as positive or negative), and aggregates them to its overall sentiment as either positive, neutral, or negative. Finally, Amazon uses small language models to generate summary paragraphs.[12] We should highlight one key difference between AI summarization and traditional methods, such as average ratings. Instead of extracting information equally from all reviews, the AI summarization algorithm only considers sentiments of reviews that mention the extracted keywords, while ignoring the remaining reviews. Apple deploys a similar algorithm in their App Store.[13]

To our knowledge, the existing research on AI summarization is still work-in-progress. Wang et al. (2025) and Wang and Wang (2025) find that the introduction of AI summarization increases purchase rates and shifts users from reading more reviews to exploring more listings. Su et al. (2025) study the impact of AI summarization on the content richness and the number of following user-generated reviews. To the best of our knowledge, no existing research has studied whether misinformation is overrepresented by AI summaries.

3.2 Fake Reviews

The market for fake reviews works as follows. Third-party sellers on Amazon 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.[14] These review manipulators offer consumers deals to write 5-star reviews for them in exchange for a full refund and, sometimes, an additional payment (Huang et al., 2023). Due to the high costs of buying fake reviews and the large number of third-party sellers in the same market, there are very few negative fake reviews attacking competitors on Amazon (He et al., 2022b). 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). Prior research also finds that Amazon does not engage in this type of fake review practice for products that are directly sold by Amazon (He et al., 2022b).

Some research also 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) show that fake reviews exhibit excessively strong sentiments. Luca and Zervas (2016) also have similar findings in the

reviews filtered and then removed by Yelp. Moreover, previous research has noticed that many fake reviews are highly similar to each other (e.g., Jindal and Liu, 2008; Mukherjee et al., 2012; Rayana and Akoglu, 2015; Wang et al., 2023; Gupta et al., 2024), and that fake reviewers tend to work collaboratively in groups (Rathore et al., 2021). Additionally, He et al. (2022b) notice that fake reviews are longer than authentic reviews on Amazon.

Textual features of fake reviews are consistent with how the market of fake reviews works. When reviewers have to write favorable comments for an item that they are not that familiar with, they tend to include common and general merits, making fake reviews resemble each other. In contrast, when consumers have idiosyncratic good personal experiences with the products, they tend to include more heterogeneous details. 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.

4 Theoretical Framework

In this section, we first define the overrepresentation of fake reviews. Then, we hypothesize that AI summaries overrepresent fake reviews and propose two effects of overrepresentation on information and market outcomes.

4.1 Definition of Overrepresentation

AI summaries of reviews represent consumer-generated reviews that mention the extracted keywords. Reviews that mention more of these extracted keywords appear more in AI summaries, receive greater representation in AI summaries, and are therefore more likely to shape the content and sentiment of AI summaries. We say that fake reviews are overrepresented if, on average, a fake review mentions more keywords extracted by AI summarization algorithms than an authentic one.

More formally, let W_{ij} denote the number of extracted keywords that review i of product j mentions, and let $F_{ij} \in \{0, 1\}$ indicate whether the review is fake. $F_{ij} = 1$ if review i of product j is fake and $F_{ij} = 0$ if review i of product j is authentic.

Definition 1 (Overrepresentation of Fake Reviews)**.** *We define the overrepresentation of fake reviews as a condition in which, on average, a fake review mentions more keywords extracted by AI summarization algorithms than an authentic review, i.e.,* $\mathbb{E}[W_{ij} \mid F_{ij} = 1] > \mathbb{E}[W_{ij} \mid F_{ij} = 0]$.

This definition is consistent with the literature on overrepresentation. In that literature, overrepresentation typically occurs when individuals with certain characteristics appear in a salient representation more frequently than others, thereby making the representation biased. For example, prior studies examine overrepresentation in settings such as children's books (Adukia et al., 2023), juries (Anwar et al., 2022), and public-sector employment (Gomes and Kuehn, 2025). Our definition follows the same logic: fake reviews are overrepresented if they appear in AI summaries more frequently than authentic reviews.

4.2 Intuition for Overrepresentation

As explained in Section 3.1, AI summarization is designed to summarize common themes across different user-generated reviews instead of idiosyncratic opinions.[1] As introduced in Section 3.2, there is a consensus in the literature that compared to authentic reviews, fake reviews are more similar to each other (Jindal and Liu, 2008; Mukherjee et al., 2012; Rayana and Akoglu, 2015; Wang et al., 2023; Gupta et al., 2024). Ott et al. (2011) notice that fake reviews tend to include less concrete language and are less specific than authentic reviews. Additionally, on Amazon, they are also longer (He et al., 2022b). All these findings suggest that fake reviews are likely to repeat generic and recurring themes.

Therefore, it is reasonable to hypothesize that AI summaries overrepresent fake reviews. We make Hypothesis 1, and empirically test whether the textual features of fake reviews give fake reviews more extracted keywords in AI summaries.

Hypothesis 1 (Overrepresentation of Fake Reviews)**.** *AI summaries overrepresent fake reviews, i.e.,* $\mathbb{E}[W_{ij} \mid F_{ij} = 1] > \mathbb{E}[W_{ij} \mid F_{ij} = 0]$.

4.3 Effect 1: Bias in Sentiment

We discuss the negative effects of the overrepresentation of fake reviews. According to the institutional background introduced in Section 3.2, fake reviews are deliberately asked to be positive.

On Amazon, almost all fake reviews are positive (He et al., 2022b). The overrepresentation of the overly positive portion among all reviews of review manipulators can make AI summaries of review manipulators more positive. This leads to Hypothesis 2.

Hypothesis 2 (Bias in Sentiment)**.** *Compared with other products, products with more fake reviews receive more positive sentiments corresponding to extracted keywords and more positive AI summary paragraphs.*

4.4 Effect 2: Bias in Market Outcomes

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 et al., 2025; Wang and Wang, 2025). AI summaries of reviews are positioned in a more prominent location above all user-generated reviews. Hence, AI summaries 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 summarization, 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.

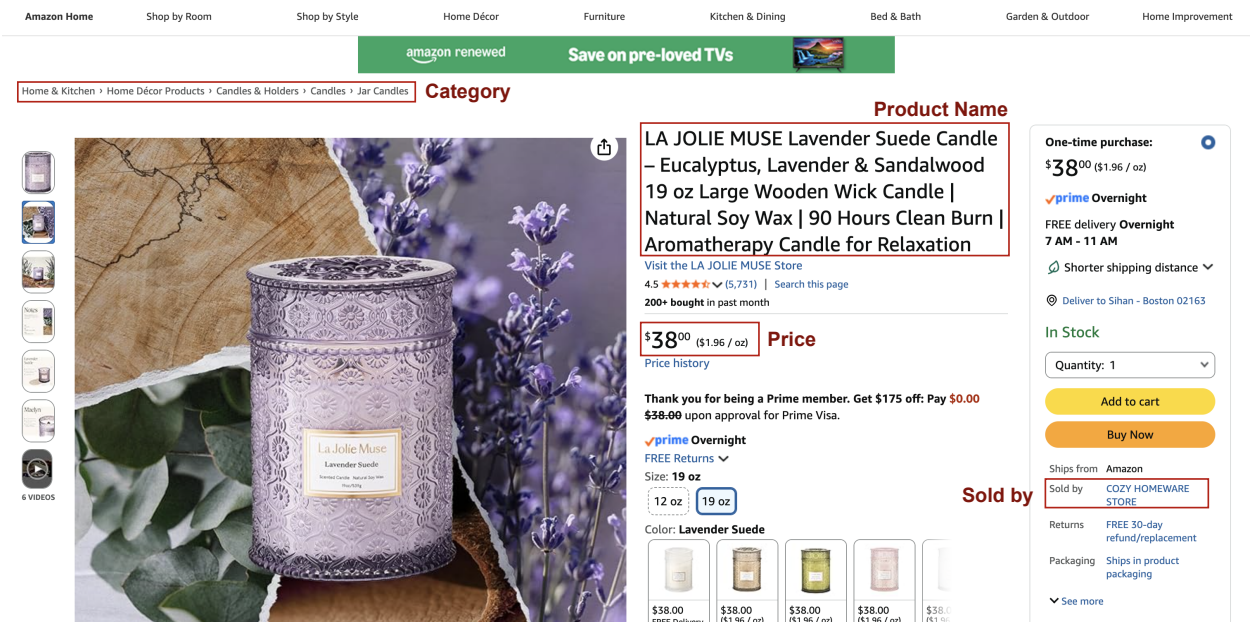
Hypothesis 3 (Bias in Market Outcomes)**.** *After the introduction of AI summarization, compared with other products, products with more fake reviews experience greater sales.*

5 Data

This paper relies on four data sources. The first is Amazon. This is our main dataset. From Amazon, we download (1) AI summary paragraphs, (2) extracted keywords with corresponding sentiments, (3) the number of mentions of each keyword (i.e., the number of reviews that mention each keyword), (4) the total number of reviews, (5) distribution of ratings (shares of 1-star, 2-star, 3-star, 4-star, and 5-star reviews), (6) average ratings, (7) product names, (8) prices of products,[15] (9) categories of products,[16] and (10) whether the product is sold by Amazon. They are all displayed on the Amazon interface. (1) - (6) can be easily observed in Figure 1, while (7) - (10) are highlighted on the Amazon interface shown in Figure 2. When collecting data, we identify the most popular search

terms in the US and around the world from reports by commercial consulting firms, including Exploding Topics[17] and Glimpse.[18] Then, we search these search terms on Amazon, and record products on the first seven pages of the results.[19] We collected 16,851 products for 159 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. A small share of products have 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 of 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.

Figure 2: Locations of Product Categories, Product Names, Product Prices, and Amazon-Sold Status on the Amazon Interface[20]



The second data source is a dataset compiled by Brett Hollenbeck from various studies, publicly available on his GitHub page.[21] It contains products predicted to have bought fake reviews by He et al. (2022a) and fake reviewers predicted by Feldman et al. (2025). We use this dataset to analyze textual features of fake reviews. In the summary statistics reported in Table 1, the number of reviews is much lower than in data source 1. We have confirmed that this is a subset of all reviews that can be classified as fake or authentic with enough confidence. We classify a review as

a fake review if it is (i) a 5-star review, (ii) written for a review manipulating seller classified by He et al. (2022a), and (iii) written by a fake reviewer classified by Feldman et al. (2025). We augment the dataset by collecting lists of keywords of the products in it from Amazon on September 29, 2025.

The third data source is RateBud.[7] We collect from RateBud a 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 2, we find that most of the products in data source 1 are of grade A (Excellent) or grade B (Very Good).

The fourth data source is Keepa.[22] 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 classic 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.

6 Overrepresentation of Fake Reviews

In this section, we test Hypothesis 1 by providing evidence that AI summaries overrepresent fake reviews. Meanwhile, we also show empirical evidence that is consistent with the intuition for the overrepresentation of fake reviews.

6.1 Empirical Strategy

Overrepresentation of fake reviews, as defined in Definition 1 (i.e., $\mathbb{E}[W_{ij} \mid F_{ij} = 1] > \mathbb{E}[W_{ij} \mid F_{ij} = 0]$), involves the joint distribution of two random variables F_{ij} (the indicator of whether review i of product j is fake) and W_{ij} (the number of extracted keywords that review i of product

Table 1: Summary Statistics for Data Sources 1 (Amazon), 2 (Fake Reviews), and 4 (Keepa)

	Mean	STD	25%	50%	75%	Non-Missing Observations
Data Source 1: Amazon						
Number of Reviews	6,815.17	24,534.75	125	859	4,242	16,851
Average Rating	4.431	0.334	4.3	4.5	4.6	16,457
Price	175.65	811.48	20.68	42.50	139.99	16,851
Amazon-Sold Product	30.08%					16,812
Data Source 2: Dataset Compiled from He et al. (2022a) and Feldman et al. (2025)						
Number of Reviews	112.64	184.41	17	46	127	3,389
Percentage of Fake Reviews	17.93%	27.23%	0	0	32.72%	3,389
Data Source 4: Keepa						
Average Sales Rank	62,439.87	282,050.98	2,892.09	12,463.65	44,653.49	11,523
Median Sales Rank	51,691.33	296,932.90	1,528	7,355	29,972	11,523

Note: This table presents the summary statistics of data sources 1, 2, and 4. In the sample we collected from data source 1, 14,479 products have AI summaries. According to Amazon, AI summaries are shown when "shared mentions" are abundant enough.[23] The missing average ratings (394 products) in the data are mainly due to no reviews, and the missing Amazon-Sold product indicators (39 products) are mainly because of no items in stock when we collected the data. In the summary statistics for data source 4 (Keepa), we first calculate average and median sales ranks across time for each product, and then report the statistics on the product level.

Table 2: Summary Statistics for Data Source 3 (RateBud)

Grade	A	B	C	D	E	F	Total
Meaning	Excellent	Very Good	Good	Fair	Poor	Caution	
Count	9,998	4,777	1,306	276	0	287	16,644
Percentage	60.07%	28.70%	7.85%	1.66%	0%	1.72%	100%

j mentions). However, we cannot directly observe F_{ij} or W_{ij}, because fake reviews are not easily observable and we do not know exactly which reviews contribute to each AI summary keyword.

6.1.1 Proxy for F_{ij}

To solve the challenge that we cannot observe F_{ij}, we use three proxies for fake reviews (review i of product j with $F_{ij} = 1$) or products with more fake reviews (product j with higher $f_j \equiv \mathbb{E}[F_{ij} \mid j]$).

1. **Amazon-sold products:** Amazon very rarely participates in review manipulation (He et al., 2022b), so we use Amazon-sold products as a proxy for products with fewer fake reviews (lower f_j).

2. **Suspicious products classified by RateBud:** We collect suspicious products from RateBud,[7] a commercial website that evaluates the credibility of reviews of products on Amazon. RateBud provides letter grades for review credibility summarized in Table 2. A and B are largely free from fraud, while D, E, and F are suspected manipulators. C, in the middle, is somewhat ambiguous. In the main text, we regard products in C, D, E, and F as a subgroup of products with more fake reviews (higher f_j). We report the results when we regard products in D, E, and F as a subgroup of products with more fake reviews (higher f_j) in Appendix A.2. The results are largely consistent.

3. **Fake reviews in the literature:** In Data Source 2, we classify a review as a fake review ($F_{ij} = 1$) if it is (i) a 5-star review, (ii) written for a review manipulating seller classified by He et al. (2022a), and (iii) written by a fake reviewer classified by Feldman et al. (2025).

6.1.2 Proxy for W_{ij}

To solve the challenge that we cannot observe W_{ij}, we consider two methods. First, we run benchmark algorithms (elaborated in Section 6.2) by ourselves to see how many extracted keywords each review mentions. Then for each review in Data Source 2,[21] we have a proxy for both W_{ij} and F_{ij}, allowing us to test the overrepresentation of fake reviews based on the empirical joint distribution of W_{ij} and F_{ij}.

Second, we turn to the product-level analysis that is feasible with the data we collected from Amazon. We can estimate $w_j \equiv \mathbb{E}[W_{ij}|j]$, which is the mean of W_{ij} for product j, using the sample

average $\hat{w}_j$, defined as the summation of the number of reviews that mention each keyword (i.e., the total number of mentions of all keywords) divided by the total number of reviews. They can be directly observed on the Amazon interface.

However, in order to use the results from the product-level analysis (joint distribution of w_j and f_j) to test the overrepresentation hypothesis (Hypothesis 1) that concerns the joint distribution of W_{ij} and F_{ij}, we need to make Assumption 1, which states that the number of keywords mentioned by reviews of different products have the same mean conditional on whether they are fake or authentic. We also empirically test the plausibility of Assumption 1 in Section 6.3.2, in which we study how $\mathbb{E}[W_{ij} \mid F_{ij} = 0, j]$ varies across j.

Assumption 1 (Mean Independent across Products)**.** $\mathbb{E}[W_{ij} \mid F_{ij}, j] = \mathbb{E}[W_{ij} \mid F_{ij}]$.

Now, consider two subgroups of products J_1, J_2. Products in J_1 have more fake reviews than those in J_2. Formally, $\mathbb{E}[f_j \mid j \in J_1] > \mathbb{E}[f_j \mid j \in J_2]$.

Proposition 1. *Under Assumption 1,* $\mathbb{E}[W_{ij} \mid F_{ij} = 1] > \mathbb{E}[W_{ij} \mid F_{ij} = 0]$ *if and only if* $\mathbb{E}[w_j \mid j \in J_1] > \mathbb{E}[w_j \mid j \in J_2]$.

Proof. For an arbitrary subgroup of products J,

$$\begin{aligned}
\mathbb{E}[w_j \mid j \in J] =& \mathbb{E}[\mathbb{E}[W_{ij} \mid j] \mid j \in J] \\
=& \mathbb{E}[f_j \mathbb{E}[W_{ij} \mid F_{ij} = 1, j] + (1 - f_j)\mathbb{E}[W_{ij} \mid F_{ij} = 0, j] \mid j \in J] \\
=& \mathbb{E}[f_j \mathbb{E}[W_{ij} \mid F_{ij} = 1] + (1 - f_j)\mathbb{E}[W_{ij} \mid F_{ij} = 0] \mid j \in J] \\
=& \mathbb{E}[f_j \mid j \in J]\mathbb{E}[W_{ij} \mid F_{ij} = 1] + (1 - \mathbb{E}[f_j \mid j \in J])\mathbb{E}[W_{ij} \mid F_{ij} = 0] \\
=& \mathbb{E}[f_j \mid j \in J](\mathbb{E}[W_{ij} \mid F_{ij} = 1] - \mathbb{E}[W_{ij} \mid F_{ij} = 0]) + \mathbb{E}[W_{ij} \mid F_{ij} = 0]
\end{aligned}$$

The first equality is due to the definition of w_j, the second equality is due to the law of total probability, the third equality follows from Assumption 1, the fourth equality takes items independent of j outside the expectation, and the fifth equality is a rearrangement of terms. Then,

$$\mathbb{E}[w_j \mid j \in J_1] - \mathbb{E}[w_j \mid j \in J_2] = (\mathbb{E}[f_j \mid j \in J_1] - \mathbb{E}[f_j \mid j \in J_2])(\mathbb{E}[W_{ij} \mid F_{ij} = 1] - \mathbb{E}[W_{ij} \mid F_{ij} = 0])$$

Because $\mathbb{E}[f_j \mid j \in J_1] > \mathbb{E}[f_j \mid j \in J_2]$, $\mathbb{E}[W_{ij} \mid F_{ij} = 1] > \mathbb{E}[W_{ij} \mid F_{ij} = 0]$ if and only if

$\mathbb{E}[w_j \mid j \in J_1] > \mathbb{E}[w_j \mid j \in J_2]$. □

Proposition 1 allows us to test overrepresentation $\mathbb{E}[W_{ij} \mid F_{ij} = 1] > \mathbb{E}[W_{ij} \mid F_{ij} = 0]$ that involves the joint distribution of two unobservable variables by comparing $\hat{w}_j$ between subgroups of products with more fake reviews (higher f_j) and fewer fake reviews (lower f_j).

6.2 Review-Level Analysis

We first directly approximate $\mathbb{E}[W_{ij} \mid F_{ij}]$ by running widely used algorithms that can output W_{ij}. In Data Source 2,[21] we already have an approximation for F_{ij}. Therefore, we can directly test the overrepresentation of fake reviews without further assumptions.

6.2.1 Textual Features of Reviews

We first calculate the basic textual features of reviews including the length, sentiment, and textual similarities. We select products that have both fake reviews and authentic reviews. 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,[24] and set the temperature (i.e., the randomness in generating the results) to zero, so that the results are replicable.[25] To compare similarities, we first convert reviews into embeddings using the benchmark model all-MiniLM-L6-v2, and then calculate the cosine similarities of the embeddings of reviews within the same product. We also try other embedding models as a robustness check in Appendix B.1.

The results are shown in Panel A of Table 3. The p-values are from Welch's t-tests using fake reviews as the reference group. 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 regardless of whether we compare fake reviews with all authentic reviews or authentic 5-star reviews. These results replicate the findings in the literature and support the intuition for the overrepresentation of fake reviews discussed in Section 4.2.

6.2.2 Comparing the Number of Extracted Keywords

We test Hypothesis 1 here. We take two steps to construct W_{ij}, i.e., the number of extracted keywords mentioned by review i of product j. In the first step, we get the list of extracted keywords; in the second step, we determine how many extracted keywords in the list each review mentions.

Step 1 (Getting the list of extracted keywords). The most straightforward way is to directly use the list of keywords extracted by Amazon and displayed on its interface. We also use BERTopic (Grootendorst, 2022) and Top2Vec (Angelov, 2020) to extract keywords from the reviews in our data set. They both use UMAP to reduce the dimensionality of the embeddings (generated by all-MiniLM-L6-v2) and HDBSCAN to cluster the embeddings. After that, BERTopic uses TF-IDF to identify salient n-grams to assign as the keyword to each cluster of embeddings, while Top2Vec constructs a topic vector as the centroid of the review embeddings assigned to each cluster and then ranks candidate words and phrases by their cosine similarity to this topic vector. In each method, we retain the top representative phrase for each topic after filtering out generic marketplace terms and near-duplicate expressions, yielding a product-specific keyword set.

Step 2 (Determining how many extracted keywords each review mentions). We use two complementary criteria. First, we apply case-insensitive whole-phrase lexical matching. Second, we apply sentence-level semantic matching: each review is split into sentences, and a keyword is counted as mentioned if the cosine similarity between the keyword embedding and at least one sentence embedding in the review exceeds a fixed threshold. We use threshold 0.8 in the main text and report the analysis with other thresholds in Appendix B.2 for a robustness check. A keyword is treated as mentioned if either the lexical or semantic criterion is satisfied. Finally, we count the number of distinct extracted keywords mentioned by review i of product j, yielding three constructions of W_{ij}.

The results of the comparisons between fake reviews and authentic reviews are reported in Panel B of Table 3. The p-values come from Welch's t-tests using fake reviews as the reference group. We report all the results in Panel B using other embeddings as a robustness check in Appendix B.1. The scale varies, because TF-IDF in BERTopic selects words that are mentioned the most frequently, while Top2Vec selects words that match the centroid of embeddings. No matter which method we use, the average W_{ij} of fake reviews is significantly larger than the average W_{ij} of authentic

reviews, indicating that $\mathbb{E}[W_{ij}|F_{ij} = 1] > \mathbb{E}[W_{ij}|F_{ij} = 0]$. This supports the overrepresentation of fake reviews.

6.2.3 Robustness Check

One concern is that the way we determine whether a review mentions an extracted keyword is different from what is actually used by Amazon. We perform the same analysis with other embedding models in Appendix B.1 and other similarity thresholds in Appendix B.2 to alleviate the concern. To further address the remaining concerns, we try a third method that does not involve lexical or semantic matching to construct W_{ij}. We cut reviews into sentences, convert sentences to embeddings, use UMAP to reduce the dimensionality of the embeddings, and run HDBSCAN to cluster the embeddings of sentences. Finally, we count the number of sentences that fall into clusters for each review and use it as W_{ij}. The results are reported in Panel C of Table 3. To further alleviate this concern, we perform a product-level analysis on the output of the algorithms actually used by Amazon to determine whether a review mentions an extracted keyword. The results are reported in Section 6.3.

Another concern is that some reviews may have been removed between when Data Source 2 was collected and when the keyword list was collected. We exclude reviews that were removed by Amazon identified by comparing with the reviews documented in Hou et al. (2024). The results are reported in Panel C of Table 3. The magnitude of the difference is even larger than that reported in Panel B of Table 3.

6.3 Product-Level Analysis

We turn to product-level analysis with real algorithms deployed by Amazon. Specifically, we compare the average number of keywords represented in AI summaries ($\hat{w}_j$) between products with more (higher f_j) and fewer fake reviews (lower f_j). $\hat{w}_j$ is calculated as the total number of mentions of all keywords (i.e., the summation of the number of reviews that mention each keyword) divided by the total number of reviews. AI summary keywords and their mentions are downloaded directly from Amazon (Data Source 1 in Section 5). We use the first two proxies introduced in Section 6.1.1 to identify products with more fake reviews, i.e., non-Amazon products and products with low

Table 3: Review-Level Analysis

	Fake Reviews	Authentic Reviews	Authentic 5-Star Reviews
Panel A: Textual Features of Fake Review VS Authentic Reviews			
Length (# Characters)	302.30	230.48 ($p < 0.001$)	244.64 ($p < 0.001$)
Length (# Words)	57.66	46.02 ($p < 0.001$)	43.43 ($p < 0.001$)
Sentiment	4.72		4.66 ($p < 0.001$)
Cosine Similarity	0.4287	0.2993 ($p < 0.001$)	0.3708 ($p < 0.001$)
Panel B: Testing Hypothesis 1 by Comparing # Extracted Keywords Mentioned by Fake Reviews VS Authentic Reviews			
# Keywords Extracted by Amazon	0.4000	0.2104 ($p < 0.001$)	0.2581 ($p < 0.001$)
# Keywords Extracted by BERTopic	0.7132	0.5268 ($p < 0.001$)	0.5119 ($p < 0.001$)
# Keywords Extracted by Top2Vec	0.1765	0.0993 ($p < 0.001$)	0.1238 ($p < 0.001$)
Panel C: Robustness Check			
Using HDBSCAN to Decide Mentions	2.3273	1.4859 ($p < 0.001$)	1.5821 ($p < 0.001$)
Excluding Reviews Removed by Amazon	0.4164	0.2194 ($p < 0.001$)	0.2676 ($p < 0.001$)

Note. This table compares fake reviews with all authentic reviews and with authentic 5-star reviews in the archived review dataset. Reported p-values are from Welch's t-tests using fake reviews as the reference group. In Panel A, review length is measured by the number of characters and words in the review text; sentiment is scored by OpenAI GPT-4.1; and cosine similarity is the average within-product pairwise cosine similarity of review embeddings. Panel B tests Hypothesis 1 by comparing W_{ij}, i.e., the number of extracted keywords, mentioned by fake reviews versus authentic reviews. W_{ij} is constructed in two steps. First, for each product, we obtain a list of keywords by collecting from the interface of Amazon, or running BERTopic or Top2Vec. Second, we determine how many extracted keywords each review mentions. Panel C reports robustness checks. In the first row, mentions are determined using HDBSCAN-based clustering of sentence embeddings, and W_{ij} is defined as the number of clustered sentences in each review. In the second row, reviews later removed by Amazon are excluded from the analysis.

grades from RateBud. Here, we regard C, D, E, and F as low grades. In Appendix A.2, we report results with D, E, and F as low grades. Proposition 1 allows us to use this product-level analysis to prove the overrepresentation of fake reviews.

The results are presented in Table 4. The p-values are from Welch's t-tests. 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. We compare $\hat{w}_j$, which is the total number of mentions of all keywords (i.e., the summation of the numbers of reviews that mention each keyword) divided by the total number of reviews. For robustness, we also compare the minimum number of mentions and the average number of mentions across reviews divided by the total number of reviews. All three measures consistently show that products with more fake reviews have more keywords represented in AI summaries. Intuitively, this suggests that the reviews of fake review products contribute more to AI summaries and are therefore more represented in the sentiment of AI summaries. According to Proposition 1, we prove the overrepresentation of fake reviews.

Table 4: Product-Level Analysis

Panel A: Comparison between Amazon-Sold Products and Non-Amazon Products			
	Non-Amazon Products	Amazon-Sold Products	P-Value
Average # Mentions/# Reviews	0.0644	0.0451	<0.001
Minimum # Mentions/# Reviews	0.0339	0.0227	<0.001
Total # Mentions/# Reviews ($\hat{w}_j$)	0.4540	0.3281	<0.001
Panel B: Comparison between Predicted Fake Review Products by RateBud and Other Products			
	Fake Review Products	Other Products	P-Value
Average # Mentions/# Reviews	0.1231	0.0550	<0.001
Minimum # Mentions/# Reviews	0.0780	0.0280	<0.001
Total # Mentions/# Reviews ($\hat{w}_j$)	0.7482	0.3974	<0.001
Panel C: Comparison Predicted Fake Review Products and Other Non-Amazon Products			
	Fake Review Products	Other Non-Amazon Products	P-Value
Average # Mentions/# Reviews	0.1233	0.0610	<0.001
Minimum # Mentions/# Reviews	0.0777	0.0313	<0.001
Total # Mentions/# Reviews ($\hat{w}_j$)	0.7557	0.4370	<0.001

Note: This table tests Hypothesis 1 by comparing the average number of keywords represented in AI summaries ($\hat{w}_j$) between products with more versus fewer fake reviews. Reported p-values are from Welch's t-tests. For robustness, we also compare the average number of mentions across keywords and the minimum number of mentions across keywords divided by the total number of reviews. In the theoretical framework in Section 4, this proves the existence of the overrepresentation of fake reviews.

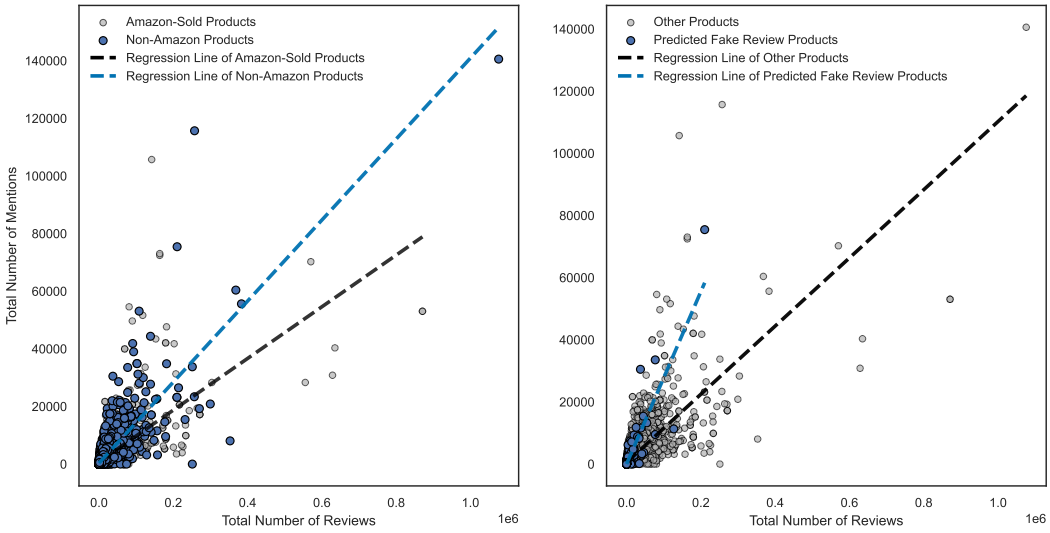
6.3.1 Robustness Check

One concern with this analysis is that there may be an upper bound on the number of keywords that can be displayed on the Amazon interface. Products with fewer fake reviews may be more popular with a larger number of reviews, but their number of keywords is bounded. Therefore, some of their keywords are dropped, which can explain why on average their reviews mention fewer keywords. In Appendix A.1, we report the results on the subset of products that are not bounded by the upper bound. The results are largely consistent with Table 4, suggesting that this explanation is at least not the whole story.

Another concern is that the results in Table 4 are driven by the fact that we normalize the metrics by dividing the total number of reviews. To mitigate this concern, in Figure 3, we plot the total number of mentions of all keywords against the total number of reviews. It is clear in Figure 3 that the majority of products with more fake reviews are above the regression line of products with fewer fake reviews. Additionally, the regression line for products with more fake reviews is steeper than and is above the regression line for products with fewer fake reviews. This result shows that no matter what the total number of reviews is, AI summarization systematically takes

more reviews into sentiment analysis from the reviews of fake review products.

Figure 3: Evidence for Overrepresentation of Fake Reviews without Normalization



Note: This figure plots the total number of mentions of all keywords, against the total number of reviews. The figure shows that for products with any total number of reviews, AI summarization algorithms take more reviews of fake review products into sentiment analysis and represent them in the sentiments of AI summaries.

A further concern is that Amazon may only select a fixed number of reviews to extract keywords from. However, it is evident from Figure 3 that the number of reviews that Amazon uses to extract keywords and analyze sentiments are constantly increasing with the total number of reviews without reaching an upper bar.

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.3. The results are consistent with the findings in Table 4.

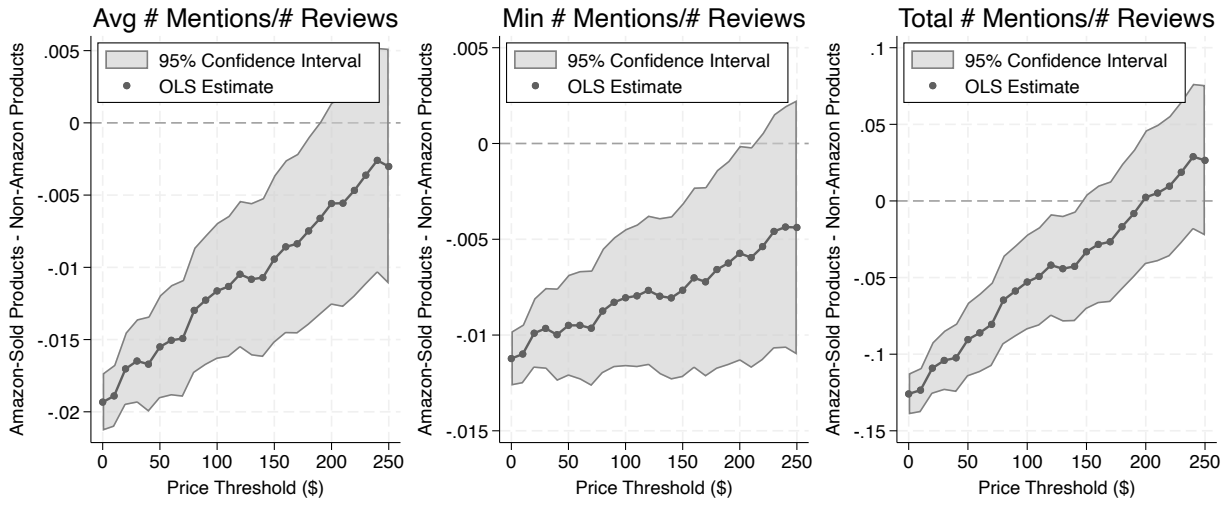
6.3.2 Plausibility of the Identification Assumption

Here, we test the plausibility of Assumption 1, which is the bridge that connects the product-level evidence to the overrepresentation of fake reviews. We would like to study how much $\mathbb{E}[W_{ij} \mid F_{ij} = 0, j]$ varies across different products j. We take advantage of the nature of the market for fake reviews to conduct a falsification test that can shed light on the plausibility of the assumption 1. Amazon is a platform with relatively strict 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 et al., 2022b). 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 have very few fake reviews, implying that $w_j \approx \mathbb{E}[W_{ij} \mid F_{ij} = 0, j]$. If Assumption 1 holds, there should be no significant differences in $w_j \approx \mathbb{E}[W_{ij} \mid F_{ij} = 0, j]$ between subgroups of products.

We plot the difference (Amazon-Sold Products - Non-Amazon Products) in average # mentions/# reviews, minimum # mentions/# reviews, and total # mentions/# reviews ($\hat{w}_j$) in Panel A of Table 4 on the subsample of products with prices higher than a threshold. The results are reported in Figure 4. It is clear from the figure that the difference in the total, minimum, and average number of mentions divided by the total number of reviews is only significant in magnitude and statistical tests only when the products are fairly cheap, suggesting that Assumption 1 is plausible. The heterogeneity between products conditional on F_{ij} per se is not enough to generate the findings in Table 4. The findings are likely due to the underlying overrepresentation of fake reviews (i.e., the variation with F_{ij}).

Figure 4: Falsification Test for the Overrepresentation of Fake Reviews



7 Effects of Overrepresentation

7.1 Effect 1: Bias in Sentiment

We test Hypothesis 2, which states that the sentiment of AI summaries is biased in favor of review manipulators. Specifically, we expect that products with more fake reviews have: (i) more positive sentiments corresponding to extracted keywords, and (ii) more positive AI summary paragraphs.

7.1.1 Empirical Strategy

The general idea of the empirical analysis is straightforward. We want to compare the two dependent variables between products with more (higher $f_j \equiv \mathbb{E}[F_{ij} \mid j]$) and fewer fake reviews (lower f_j): sentiments associated with keywords and sentiments of AI summary paragraphs.

We continue to use the two proxies introduced in Section 6.1.1 to identify products with more fake reviews, i.e., non-Amazon products and products with low grades from RateBud. 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 (forbidden sign), and 0 to neutral sentiment (no sign or a short gray line). Then, we take the average across different keywords. To quantify the sentiments of AI summary paragraphs, we use two language models: OpenAI GPT-4.1[24] and bert-base-multilingual-uncased-sentiment.[26] When using OpenAI GPT-4.1, we prompt the model to predict the quality perceived by consumers who see the summary paragraph, on the scale of average ratings.[27] bert-base-multilingual-uncased-sentiment is a specialized language model finetuned on the task of predicting the ratings of reviews. We report the results with OpenAI GPT-4.1 in the main text, and the results with bert-base-multilingual-uncased-sentiment in Appendix C.1.

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

$$DV_j = \beta I_j + Rating_Controls_j \gamma + \phi_c + \psi_{kn} + \epsilon_j \tag{1}$$

where j denotes products, c denotes categories, k denotes search terms, and n denotes the page number in the search results on which product j is shown. DV_j is one of the two dependent

variables that we just mentioned, including sentiments corresponding to keywords and sentiments of AI summary paragraphs. I_j is the indicator for products with more (or less) fake reviews based on the three proxies discussed in Section 6.1.1. $Rating_Controls_j$ is a vector of flexible controls of the rating distribution. ϕ_c is category fixed effects and ψ_{kn} is search term $\times$ 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.

7.1.2 Results

The results are shown in Table 5. We consider I_j to indicate whether product j is sold by Amazon in Columns (1)-(4) and to indicate whether product j is a fake review product predicted by RateBud in Columns (5)-(8). We group C, D, E, and F together as fake review products in Table 5. In Appendix C.3, we report the results with only D, E and F as fake review products as a robustness check.

The results of the sentiments corresponding to keywords are reported in Columns (1), (2), (5) and (6), while the results of the sentiments of AI summary paragraphs are reported in Columns (3), (4), (7) and (8). We find that with search term $\times$ 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, and significantly more negative sentiments of summary paragraphs than the other products on the same search result page. The results on the classified fake review products are exactly the opposite. R^2 of the regressions is considerable, suggesting that we have successfully controlled for many factors that can influence the dependent variables. The results validate Hypothesis 2.

7.1.3 Robustness Check

We present two robustness checks in Table 6. First, we put the two group indicators, predicted fake review products and Amazon-sold products, together in one regression. Columns (1) - (4) report the results. Second, we continue the stricter comparison in Panel C of Table 4, in which we compare fake review products classified by RateBud and the other products within non-Amazon products. Columns (5) - (8) report the results. Both robustness checks yield results consistent with

Table 5: Bias in Sentiment

	(1) Keyword	(2) Keyword	(3) Paragraph	(4) Paragraph	(5) Keyword	(6) Keyword	(7) Paragraph	(8) Paragraph
Amazon-Sold	-0.0469***	-0.0450***	-0.0861***	-0.0807***				
	(0.00475)	(0.00476)	(0.00882)	(0.00883)				
Fake					0.0440***	0.0473***	0.116***	0.112***
					(0.00900)	(0.00902)	(0.0167)	(0.0167)
Avg Rating	0.692***	0.835***	1.165***	1.047***	0.694***	0.864***	1.172***	1.121***
	(0.00846)	(0.0393)	(0.0157)	(0.0729)	(0.00854)	(0.0399)	(0.0158)	(0.0740)
# Reviews		-2.36e-07**		-6.64e-07***		-3.25e-07***		-8.40e-07***
		(7.96e-08)		(1.48e-07)		(8.14e-08)		(1.51e-07)
Variance of Ratings		0.0642***		0.0526***		0.0719***		0.0697***
		(0.00849)		(0.0158)		(0.00864)		(0.0160)
Share of 5-Star Reviews		-0.0271		0.705***		-0.0583		0.615***
		(0.0780)		(0.145)		(0.0791)		(0.146)
Constant	-2.366***	-3.067***	-0.973***	-1.045***	-2.392***	-3.202***	-1.036***	-1.358***
	(0.0376)	(0.135)	(0.0697)	(0.250)	(0.0379)	(0.137)	(0.0702)	(0.254)
Search Term × Page No. FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	13515	13515	13515	13515	13382	13382	13382	13382
R^2	0.553	0.557	0.528	0.532	0.552	0.556	0.528	0.533

Note. This table tests Hypothesis 2 by comparing sentiments of AI summaries between products more versus less likely to have fake reviews. The dependent variables are (i) Keyword Sentiment, constructed by coding each keyword's sentiment as 1 (positive/checkmark), 0 (neutral/no sign), or -1 (negative/minus) and averaging across keywords, and (ii) Paragraph Sentiment, measured by prompting OpenAI GPT-4.1 to predict the perceived product quality from the AI summary paragraph. As for the proxy for products more and less likely to have fake reviews, Columns (1)-(4) use an Amazon-Sold indicator; Columns (5)-(8) use a Fake indicator where products with RateBud grades C/D/E/F are grouped as "fake review products".
Reported standard errors are in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5.

7.1.4 Discussion on the Causal Relationship

Hypothesis 2 is not a causal claim, but we want to shed some light on whether the difference in the sentiment of AI summaries is caused by the difference in the portion of fake reviews between the subgroups rather than other differences between subgroups with more and fewer fake reviews. We use three approaches to provide suggestive evidence of a causal relationship. First, we flexibly control the distribution of ratings in the specification in Equation (1). Second, we control the characteristics of products by fixed effects in Equation (1), 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 also conduct a falsification test here.

The cost of buying fake reviews increases with the price of the products because it is common for sellers to reimburse the price of the product to consumers for writing a fake review. Therefore,

Table 6: Robustness Check for Bias in Sentiment

	(1) Keyword	(2) Keyword	(3) Paragraph	(4) Paragraph	(5) Keyword	(6) Keyword	(7) Paragraph	(8) Paragraph
Amazon-Sold	-0.0463***	-0.0443***	-0.0855***	-0.0800***				
	(0.00478)	(0.00478)	(0.00886)	(0.00886)				
Fake	0.0420***	0.0456***	0.112***	0.109***	0.0343***	0.0382***	0.0978***	0.0958***
	(0.00897)	(0.00898)	(0.0166)	(0.0166)	(0.0103)	(0.0103)	(0.0190)	(0.0191)
Avg Rating	0.696***	0.871***	1.175***	1.132***	0.659***	0.897***	1.120***	1.153***
	(0.00852)	(0.0398)	(0.0158)	(0.0738)	(0.0105)	(0.0503)	(0.0195)	(0.0931)
# Reviews		-2.59e-07**		-7.18e-07***		-5.20e-07***		-1.03e-06***
		(8.14e-08)		(1.51e-07)		(1.27e-07)		(2.36e-07)
Variance of Ratings		0.0722***		0.0703***		0.0812***		0.0850***
		(0.00861)		(0.0160)		(0.0109)		(0.0202)
Share of 5-Star Reviews		-0.0727		0.589***		-0.200*		0.452*
		(0.0788)		(0.146)		(0.0966)		(0.179)
Constant	-2.386***	-3.208***	-1.020***	-1.365***	-2.218***	-3.240***	-0.779***	-1.376***
	(0.0378)	(0.137)	(0.0701)	(0.253)	(0.0466)	(0.174)	(0.0864)	(0.323)
Search Term × Page No. FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	13360	13360	13360	13360	9126	9126	9126	9126
R^2	0.556	0.560	0.532	0.537	0.565	0.570	0.542	0.546

Standard errors in parentheses

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

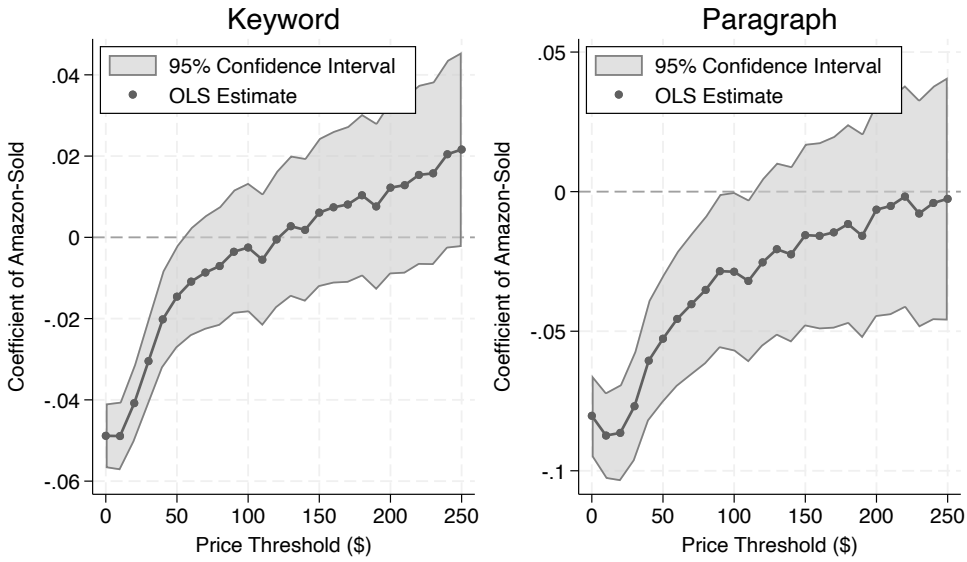
we should expect that more expensive products are less likely to have fake reviews, and that there is no significant difference in fake reviews between expensive Amazon-sold products and expensive non-Amazon products. If the difference in fake reviews plays an important role in the results, the results should no longer hold without the difference in fake reviews.

We rerun the specifications in Columns (2) and (4) of Table 5,[28] 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 5. As is evident, the coefficients are significantly negative only when the prices of the products are fairly low. In the subsample with products more expensive than $100 and thus fake reviews are very unlikely, the effects we have discovered in Table 5 no longer exist, suggesting that fake reviews are likely to be the reason underlying the difference in the two dependent variables between Amazon-sold products and non-Amazon products.

7.2 Effect 2: Bias in Market Outcomes

We test Hypothesis 3 to investigate whether, after the introduction of AI summarization, review manipulators experience greater sales than the other sellers.

Figure 5: Falsification Test for Bias in Sentiments



7.2.1 Empirical Strategy

Following the classic paper in the literature on online reviews Chevalier and Mayzlin (2006), we use the logarithms of 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 summarization. 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 j indexes products, and t indexes date. I_j is again the indicator for subgroups of products with more (higher $f_j \equiv \mathbb{E}[F_{ij} \mid j]$) or fewer fake reviews (lower f_j). We continue to use the two proxies introduced in Section 6.1.1 to identify products with more fake reviews, i.e., non-Amazon products and products with low grades from RateBud. To make the subgroups of products comparable, we run exact matching on categories and the interaction of search terms and page numbers. Specifically, we first define the matching strata as category $\times$ (keyword $\times$ page). We only compare products that are in the same product category[29] and appear on the same search result page (the same search term and the same search result page number). We then implement exact matching within these strata and drop observations in strata that contain products from only one group. The matching procedure produces a set of weights: products in one group receive weight one, while products in the other group are reweighted so that, within each stratum, the sum of weights is identical across the

two groups. Finally, we run weighted regressions using the weights generated in exact matching. Following the suggestions of Bertrand et al. (2004), we cluster the error term at the product level to correct for serial autocorrelation over time.

$$\log(Sales_Ranks_{jt}) = \beta I_j \times After_t + \alpha_j + \omega_t + \epsilon_{jt} \tag{2}$$

We also conduct event studies in Equation (3). Specifically, we replace $After_t$ with a dummy variable ϕ_m for each month 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.

$$\log(Sales_Ranks_{jt}) = \sum_m \beta_m I_j \times \phi_m + \alpha_j + \omega_t + \epsilon_{jt} \tag{3}$$

7.2.2 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% drop in sales ranks after the introduction of AI summarization. The results of the robustness check with Amazon share are reported in Appendix D.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 D.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 the classified fake reviews products improved by approximately 20% after the introduction of AI summarization. We notice that the coefficients for classified fake review products are larger and more significant after matching.

We report the results of event studies. We plot the coefficients of the interactions of each month dummy and the group indicator. Following common practice, we normalize the month when AI summarization 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 D.3. The result of the comparison between Amazon-sold and non-Amazon products is shown in

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

	(1) Log(Rank)	(2) Log(Rank)	(3) Log(Rank)	(4) Log(Rank)
Amazon-Sold × After	0.119***	0.109*		
	(0.0288)	(0.0470)		
Fake × After			-0.181*	-0.231**
			(0.0749)	(0.0840)
Constant	8.518***	8.465***	8.544***	9.110***
	(0.00560)	(0.0106)	(0.00116)	(0.00387)
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	6630267	5053113	6630267	2173630
R^2	0.809	0.816	0.809	0.813

Note. This table tests Hypothesis 3 by estimating a DiD model that compares the change in log of sales ranks after the introduction of AI summarization between products more versus less likely to have fake reviews. As for the proxy for products more and less likely to have fake reviews, Columns (1) and (2) use an Amazon-Sold indicator; Columns (3) and (4) use a Fake indicator where products with RateBud grades C/D/E/F are grouped as "fake review products". We cluster all standard errors at the product level.
Reported standard errors are in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

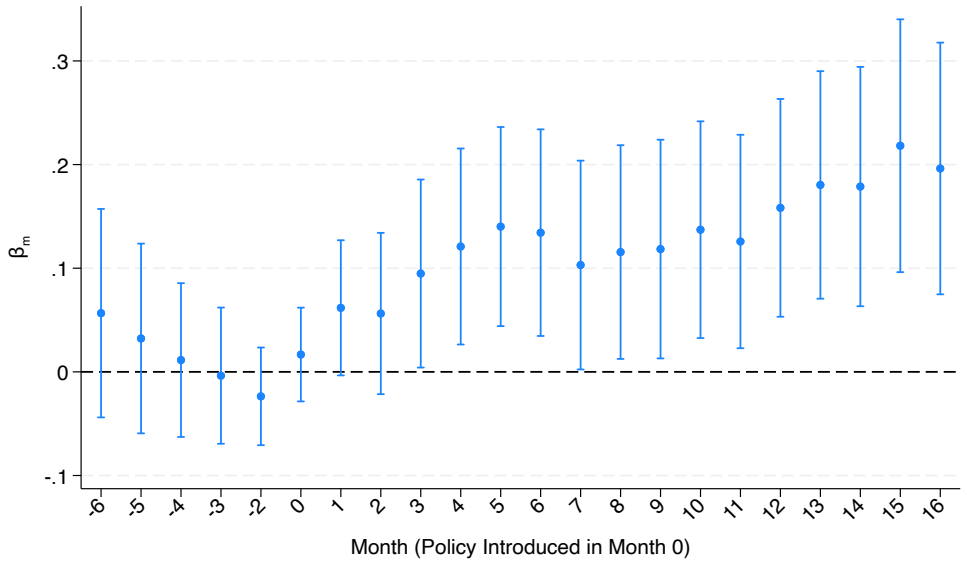
Figure 6, while the result of the comparison between predicted fake review products and other products is shown in Figure 7. There were no significant pretrends before the introduction of AI summarization. Instead, there was a significant break in the evolution of β_m after the introduction of AI summarization. The sales ranks of products with more fake reviews relatively decreased (i.e., moved up), suggesting that their sales relatively increased following the introduction of AI summarization. In Appendix D.3, we find consistent results even if we do not match the subgroups on categories and search terms.

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

8 Conclusion

Fake reviews have been shown to harm consumer welfare (He et al., 2022b; Gandhi et al., 2024), and have received much attention from the Federal Trade Commission (FTC) of the US[5] and the Parliament of the UK.[6] In this paper, we find evidence that the negative effects of fake reviews can be amplified by AI summarization of reviews, which has been a common practice on digital

Figure 6: Event Study for the Comparison between Amazon-Sold and Other Products



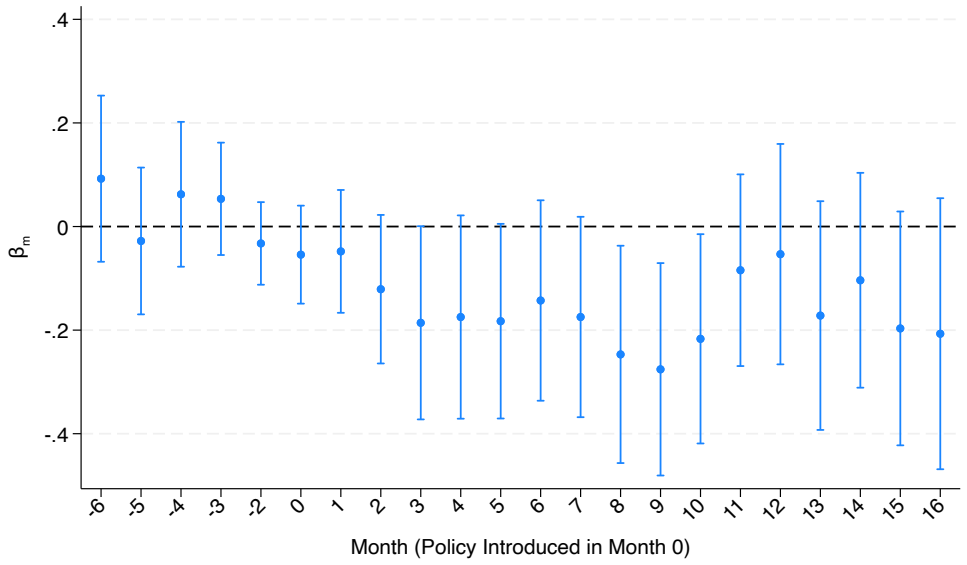
Note: The y-axis is the estimated coefficients β_m of the regression $\log(Sales_Ranks_{jt}) = \sum_m \beta_m I_j \times \phi_m + \alpha_j + \omega_d + \epsilon_{jt}$, where ϕ_m is month dummy, I_j is the group indicator of whether the product is sold by Amazon (and therefore with fewer fake reviews), α_j is product dummy, and ω_d is date dummy. We plot point estimate and 95% confidence intervals. The error terms are clustered at the product level. The x-axis is m (month). This figure shows that after the introduction of AI summarization, products that are largely manipulation-free experience a relative decrease in sales.

platforms. This AI bias is fundamentally rooted in the design of AI summarization algorithms. AI summarization algorithms are designed to summarize common themes among user-generated reviews, while fake reviews disproportionately concentrate on common themes

We first define the overrepresentation of fake reviews. In the empirical exercise, we confirm that AI summaries overrepresent fake reviews by review-level analysis and product-level analysis. Next, we empirically test two effects of overrepresentation. First, we show that the sentiments associated with keywords and AI summary paragraphs are more positive for products with more fake reviews, even after we control for flexible metrics of reviews and abundant fixed effects. Second, we find that the introduction of AI summarization benefited review manipulators more than the other sellers.

Although our empirical analyses focus on the popular e-commerce platform, Amazon, we believe our findings are generalizable to other digital platforms that also deploy AI summarization that extracts common themes from consumer reviews. As highlighted in Section 3, Apple App Store also selectively extracts opinions only from reviews that mention popular topics.[13] Because fake reviews rely heavily on repetitive templates or coordinated campaigns instead of idiosyncratic

Figure 7: Event Study for the Comparison between Predicted Fake and Other Products



Note: The y-axis is the estimated coefficients β_m of the regression $\log(Sales_Ranks_{jt}) = \sum_m \beta_m I_j \times \phi_m + \alpha_j + \omega_d + \epsilon_{it}$, where ϕ_m is month dummy, I_j is the group indicator of whether the product is graded by RateBud as C/D/E/F, α_j is product dummy, and ω_d is date dummy. We plot point estimate and 95% confidence intervals. The error terms are clustered at the product level. The x-axis is m (month). This figure shows that after the introduction of AI summarization, products with more fake reviews experience a relative increase in sales.

personal experiences, they naturally form clusters around common themes. This pattern should exist across contexts. As long as AI summarization prioritizes extracting salient and frequent themes, it is likely to remain biased toward fake reviews. We leave the investigation of the overrepresentation of fake reviews on other platforms to future research.

This distortive effect on the market is particularly concerning to society. Researchers have shown that AI summaries can increase purchase rates (Wang et al., 2025) and hence benefit platforms. This can explain why platforms are increasingly adopting AI summarization. However, our findings suggest that AI summaries might direct consumers to suboptimal products due to the interaction between the design of the 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 summarization overrepresents 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 these AI tools are still evolving, and their behavior can be difficult to anticipate. They may lead to unexpected and unintended outcomes that drive customers away.

It is challenging to design AI summarization 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 researchers in different fields to work towards algorithms proficient in summarizing information contaminated by misinformation, such as online reviews.

Notes

[1] https://www.aboutamazon.com/news/amazon-ai/amazon-improves-customer-reviews-with-generative-ai
[2] https://www.wsj.com/tech/personal-tech/fake-reviews-and-inflated-ratings-are-still-a-problem-for-amazon-11623587313
[3] https://apnews.com/article/fake-online-reviews-generative-ai-40f5000346b1894a778434ba295a0496
[4] https://www.wsj.com/us-news/youre-probably-falling-for-fake-product-reviews-b4d07f23
[5] https://www.ftc.gov/news-events/news/press-releases/2024/08/federal-trade-commission-announces-final-rule-banning-fake-reviews-testimonials
[6] https://www.legislation.gov.uk/ukpga/2024/13/contents
[7] https://www.ratebud.ai/
[8] Screenshot captured from https://www.amazon.com/dp/B0DWJCK9LZ/ on November 1, 2025
[9] Screenshot captured 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
[10] https://www.emarketer.com/content/5-charts-search-2024-google-ai-retail-media
[11] https://aws.amazon.com/blogs/machine-learning/going-beyond-ai-assistants-examples-from-amazon-com-reinventing-industries-with-generative-ai/
[12] https://www.channelmax.net/article/amazon-introduces-ai-generated-review-summaries-to-help-customers-shop-faster-and-smarter
[13] https://machinelearning.apple.com/research/app-store-review
[14] https://nymag.com/intelligencer/2022/07/amazon-fake-reviews-can-they-be-stopped.html
[15] We download the price of the default size and color.
[16] Amazon interface displays a multi-level category. When we control categories using fixed effects, we control the finest level of categories, so that we control enough unobservable confounders. When we conduct exact matching on categories, we match on the first-level category, so that we do not have to drop too many observations.
[17] https://explodingtopics.com/blog/most-searched-items-on-amazon
[18] https://meetglimpse.com/top-searched/most-searched-products-on-amazon/
[19] We also stop collecting data when we have already collected more than 110 products but not reached page seven.
[20] Screenshot captured from https://www.amazon.com/dp/B0B74R1VKS/ on March 3, 2026
[21] https://github.com/bretthollenbeck/fake-reviews-data
[22] https://keepa.com/
[23] https://www.aboutamazon.com/news/retail/amazon-ai-generated-review-highlights
[24] https://openai.com/index/gpt-4-1/
[25] The prompt we use is as follows. System prompt: "You are an AI assistant that scores the sentiment of reviews on Amazon. Respond succinctly in the requested format." User prompt: "Analyze the sentiment of reviews. What is the quality perceived by consumers? Please output a float number: X. X should be between 1 and 5. Review Text: {Review Text}"
[26] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment
[27] The prompt we use is as follows. System prompt: "You are an AI assistant that scores the sentiment of AI summary of online reviews on Amazon. Respond succinctly in the requested format." User prompt: "Analyze the sentiment of AI summary of reviews. What is the quality perceived by consumers? Please output a float number: X. X should be between 1 and 5. AI Summary: {The content of AI summary}"
[28] We omit fixed effects in subsample analysis.
[29] Here, we only consider the first-level category to avoid dropping too many observations.

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.

Adukia, A., Eble, A., Harrison, E., Runesha, H. B., and Szasz, T. (2023). What we teach about race and gender: Representation in images and text of children's books. *The Quarterly Journal of Economics*, 138(4):2225–2285.

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Anwar, S., Bayer, P., and Hjalmarsson, R. (2022). Unequal jury representation and its consequences. *American Economic Review: Insights*, 4(2):159–174.

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*.

Gomes, P. and Kuehn, Z. (2025). You're the one that i want! understanding the over-representation of women in the public sector. *American Economic Journal: Macroeconomics*.

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.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Gupta, R., Jindal, V., and Kashyap, I. (2024). Recent state-of-the-art of fake review detection: a comprehensive review. *The Knowledge Engineering Review*, 39:e8.

He, S., Hollenbeck, B., Overgoor, G., Proserpio, D., and Tosyali, A. (2022a). 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. (2022b). 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*.

Huang, Y., Villas-Boas, J. M., and Zhao, M. (2023). Unmasking the deception: The interplay between fake reviews, rating dispersion, and consumer demand. *Working Paper*.

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.

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.

Rathore, P., Soni, J., Prabakar, N., Palaniswami, M., and Santi, P. (2021). Identifying groups of fake reviewers using a semisupervised approach. *IEEE Transactions on Computational Social Systems*, 8(6):1369–1378.

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.

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.

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, H. and Wang, T. (2025). Does ai-generated review summarization affect consumer purchasing behavior?—an empirical study based on the amazon platform. In *Proceedings of the 58th Hawaii international conference on system sciences*.

Wang, J., Chen, J., and Zhang, W. (2023). A novel approach for fake review detection based on reviewing behavior and bert fused with cosine similarity. In *International Symposium on Knowledge and Systems Sciences*, pages 18–32. Springer.

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 Overrepresentation of Fake Reviews

A.1 Results on the Subset Not Bounded

A concern with the analysis in Table 4 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.1. The results are not only consistent with the results in Table 4, but also statistically very significant. Interestingly, after we switch to the subsample of products whose number of keywords is not bounded, the ratio between the total number of mentions and the total number of reviews 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.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 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.2, which are largely consistent with the results in Table 4.

A.3 With Controls for Categories

One may think the results shown in Table 4 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.3.

Appendix B Robustness Checks for Mechanism of Overrepresentation

B.1 Other Embedding Models

We rerun the analyses with a larger embedding model named paraphrase-mpnet-base-v2. The results are shown in Table S.4. The results are consistent with those reported in Table 3 in the main text.

B.2 Other Thresholds

When determining how many extracted keywords each review mentions, we use a semantic matching that treats a keyword as mentioned if the cosine similarity between the keyword embedding and at least one sentence embedding in the review exceeds a threshold. In the main text, we set the threshold to be 0.8. Here, we report the results with the threshold set to be 0.6. The results shown in Table S.5 are consistent with the results in Table 3 in the main text with the threshold as 0.8.

Appendix C Robustness Checks for Bias in Sentiments of AI Summary

C.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,[26] to convert summary paragraphs to sentiments measured on the scale of ratings. The results are shown in Table S.6.

C.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.[30] 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 are to have fake reviews. We replace Amazon-sold indicator with the continuous Amazon share variable. The result is presented in Table S.7. The result is largely consistent with that shown in the main text.

C.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.8. Results are still significant and consistent with what is shown in the main text, in spite of the smaller size of fake review products.

Appendix D Robustness Checks for Bias in Market Outcomes

D.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.[31] 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 are to have fake reviews. We then rerun the analysis in the main text. The results are presented in Columns (1) and (2) of Table S.9, which are largely consistent with our results in the main text. Products less likely to have fake reviews (with a greater Amazon share) experience a decline in sales after the introduction of AI summary.

D.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.9. The results are still largely consistent with those in the main text.

D.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 E Appendix Figures and Tables

Table S.1: Evidence for Overrepresentation in the Unbounded Subsample

Panel A: Comparison between Amazon-Sold Products and Non-Amazon Products			
	Non-Amazon Products	Amazon-Sold Products	P-Value
Average # Mentions/# Reviews	0.0828	0.0680	<0.001
Minimum # Mentions/# Reviews	0.0539	0.0434	<0.001
Total # Mentions/# Reviews ($\hat{w}_j$)	0.3539	0.2928	<0.001
Panel B: Comparison between Predicted Fake Review Products and Other Products			
	Fake Review Products	Other Products	P-Value
Average # Mentions/# Reviews	0.1417	0.0713	<0.001
Minimum # Mentions/# Reviews	0.1045	0.0447	<0.001
Total # Mentions/# Reviews ($\hat{w}_j$)	0.5539	0.3117	<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 # Mentions/# Reviews	0.1380	0.0751	<0.001
Minimum # Mentions/# Reviews	0.1013	0.0474	<0.001
Total # Mentions/# Reviews ($\hat{w}_j$)	0.5436	0.3276	<0.001

Table S.2: 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 # Mentions/# Reviews	0.1577	0.0577	<0.001
Minimum # Mentions/# Reviews	0.1074	0.0299	<0.001
Total # Mentions/# Reviews ($\hat{w}_j$)	0.8961	0.4113	<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 # Mentions/# Reviews	0.1512	0.0639	<0.001
Minimum # Mentions/# Reviews	0.1021	0.0335	<0.001
Total # Mentions/# Reviews ($\hat{w}_j$)	0.8856	0.4520	<0.001

Table S.3: Evidence for Overrepresentation with Controls for Categories

	(1) Average	(2) Average	(3) Min	(4) Min	(5) Total	(6) Total
Fake	0.0655***	0.0479***	0.0486***	0.0379***	0.333***	0.220***
	(0.00221)	(0.00207)	(0.00157)	(0.00155)	(0.0151)	(0.0137)
Amazon-Sold	-0.0174***	-0.0149***	-0.00975***	-0.00807***	-0.117***	-0.104***
	(0.000989)	(0.00111)	(0.000701)	(0.000827)	(0.00676)	(0.00735)
Constant	0.0608***	0.0616***	0.0312***	0.0315***	0.436***	0.444***
	(0.000574)	(0.000538)	(0.000407)	(0.000401)	(0.00392)	(0.00356)
Search Term × Page No. FE	No	Yes	No	Yes	No	Yes
Category FE	No	Yes	No	Yes	No	Yes
Observations	14259	13324	14259	13324	14259	13324
R^2	0.081	0.439	0.079	0.379	0.055	0.461

Standard errors in parentheses

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

Table S.4: Review-Level Analysis with paraphrase-mpnet-base-v2

	Fake Reviews	Authentic Reviews	Authentic 5-Star Reviews
Panel A: Textual Features of Fake Review VS Authentic Reviews			
Cosine Similarity	0.5767	0.4045 ($p < 0.001$)	0.5077 ($p < 0.001$)
Panel B: Testing Hypothesis 1 by Comparing # Extracted Keywords Mentioned by Fake Reviews VS Authentic Reviews			
# Keywords Extracted by Amazon	0.4056	0.2132 ($p < 0.001$)	0.2666 ($p < 0.001$)
# Keywords Extracted by BERTopic	0.7302	0.5292 ($p < 0.001$)	0.5538 ($p < 0.001$)
# Keywords Extracted by Top2Vec	0.1640	0.0854 ($p < 0.001$)	0.1084 ($p < 0.001$)

Table S.5: Review-Level Analysis with a Looser Threshold in Semantic Matching

	Fake Reviews	Authentic Reviews	Authentic 5-Star Reviews
# Keywords Extracted by Amazon	0.5913	0.3511 ($p < 0.001$)	0.4438 ($p < 0.001$)
# Keywords Extracted by BERTopic	0.7278	0.5374 ($p < 0.001$)	0.5270 ($p < 0.001$)
# Keywords Extracted by Top2Vec	0.3367	0.1992 ($p < 0.001$)	0.2413 ($p < 0.001$)

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

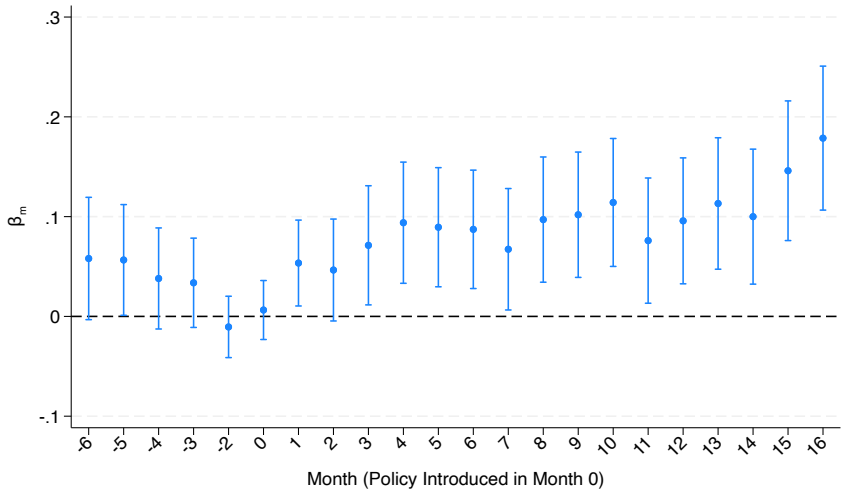


Table S.6: Bias in Sentiment with BERT

	(1) Paragraph	(2) Paragraph	(3) Paragraph	(4) Paragraph
Amazon-Sold	-0.0894***	-0.0827***		
	(0.0114)	(0.0114)		
Fake			0.173***	0.160***
			(0.0214)	(0.0215)
Avg Rating	1.194***	0.743***	1.203***	0.840***
	(0.0202)	(0.0939)	(0.0203)	(0.0953)
# Reviews		-9.28e-07***		-1.14e-06***
		(1.90e-07)		(1.94e-07)
Variance of Ratings		-0.0420*		-0.0217
		(0.0203)		(0.0206)
Share of 5-Star Reviews		1.130***		0.992***
		(0.187)		(0.189)
Constant	-1.724***	-0.502	-1.799***	-0.890**
	(0.0898)	(0.322)	(0.0903)	(0.327)
Search Term × Page No. FE	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes
Observations	13515	13515	13382	13382
R^2	0.433	0.436	0.434	0.437

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

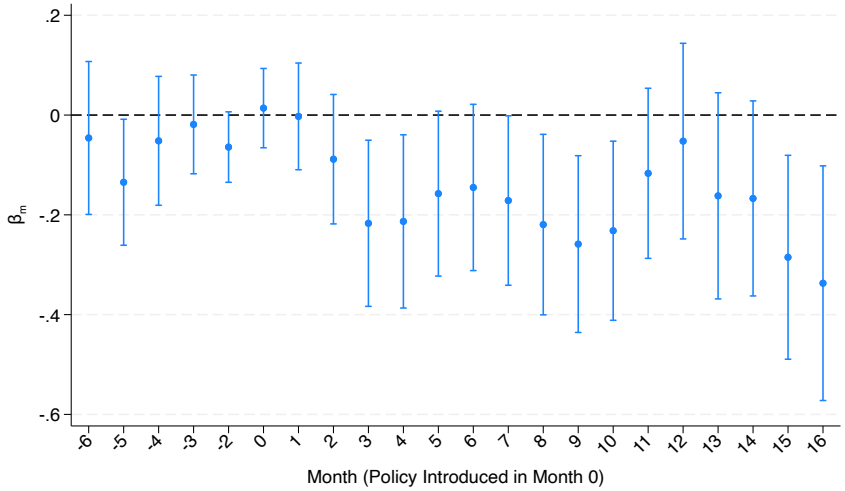


Table S.7: Bias in Sentiment with Share of Time in Buy Box

	(1) Keyword	(2) Keyword	(3) Paragraph	(4) Paragraph
amazon_share	-0.0725***	-0.0703***	-0.127***	-0.121***
	(0.00576)	(0.00577)	(0.0107)	(0.0107)
Avg Rating	0.701***	0.803***	1.181***	0.985***
	(0.00877)	(0.0407)	(0.0163)	(0.0754)
# Reviews		-1.77e-07*		-5.68e-07***
		(8.02e-08)		(1.49e-07)
Variance of Ratings		0.0588***		0.0440**
		(0.00875)		(0.0162)
Share of 5-Star Reviews		0.0621		0.891***
		(0.0820)		(0.152)
Constant	-2.404***	-2.983***	-1.039***	-0.893***
	(0.0389)	(0.139)	(0.0721)	(0.257)
Search Term × Page No. FE	Yes	Yes	Yes	Yes
Category FE	Yes	Yes	Yes	Yes
Observations	12753	12753	12753	12753
R^2	0.561	0.564	0.537	0.541

Standard errors in parentheses

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

Table S.8: Bias in Sentiment with C as Authentic

	(1) Keyword	(2) Keyword	(3) Paragraph	(4) Paragraph
Fake	0.0657*	0.0736**	0.168***	0.160**
	(0.0273)	(0.0272)	(0.0506)	(0.0505)
Avg Rating	0.675***	0.824***	1.138***	0.990***
	(0.00769)	(0.0376)	(0.0143)	(0.0696)
# Reviews		-4.79e-07***		-1.19e-06***
		(7.24e-08)		(1.34e-07)
Variance of Ratings		0.0605***		0.0386*
		(0.00821)		(0.0152)
Share of 5-Star Reviews		-0.0635		0.726***
		(0.0747)		(0.138)
Constant	-2.308***	-3.003***	-0.888***	-0.817***
	(0.0342)	(0.129)	(0.0634)	(0.239)
Search Term × Page No. FE	Yes	Yes	Yes	Yes
Observations	14315	14315	14315	14315
R^2	0.491	0.496	0.465	0.471

Standard errors in parentheses

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

Table S.9: Bias in Market Outcomes with Share of Time in Buy Box and C as Authentic

	(1) Log(Rank)	(2) Log(Rank)	(3) Log(Rank)	(4) Log(Rank)
Amazon Share × After	0.385***	0.383***		
	(0.0307)	(0.0565)		
Fake × After			0.212	0.181
			(0.266)	(0.276)
Constant	8.469***	8.403***	8.541***	9.099***
	(0.00577)	(0.0128)	(0.000407)	(0.00133)
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	6630267	5053113	6630267	2173630
R^2	0.810	0.818	0.809	0.813

Standard errors in parentheses

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