NBER WORKING PAPER SERIES

FROM COMPLAINT TO ACTION:
TECHNOLOGY-ENABLED QUALITY IMPROVEMENT FROM CONSUMER REVIEWS

Guangyu Cao
Shenghao He
Ginger Zhe Jin

Working Paper 34934
http://www.nber.org/papers/w34934

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2026, Revised July 2026

The authors are listed in alphabetical order and contribute equally to the article. We are grateful that the anonymous company that developed ARMS provided us with detailed data, offered numerous technical supports, and patiently answered many questions throughout our research. We appreciate helpful comments from Luca Bennati, seminar participants at Peking University, and conference participants at the 24th annual International Industrial Organization Conference and the 2026 Asia Meeting of the Econometric Society in China. All errors are ours. Cao acknowledges the support of the National Natural Science Foundation of China (Grant No. 72473003) and the Research Seed Fund for Young Faculty of the School of Economics, Peking University. He acknowledges the support of the National Natural Science Foundation of China (Grant No. 72192844). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

From Complaint to Action: Technology-Enabled Quality Improvement from Consumer Reviews
Guangyu Cao, Shenghao He, and Ginger Zhe Jin
NBER Working Paper No. 34934
March 2026, Revised July 2026
JEL No. D83, L15, L86, O33

ABSTRACT

This paper examines how adopting an automated review monitoring system (ARMS) allows restaurants to translate consumer feedback into operational quality improvements. ARMS provides automated alerts for negative reviews and facilitates backend work ticket management, reducing the costs of identifying, prioritizing, and addressing issues revealed by customer feedback. Using restaurant-level data on ARMS adoption and consumer reviews, we find that adoption increases average star ratings, lowers the share of negative reviews, and raises the positivity of review text. These improvements are not driven by strategic manipulation; they are greater for restaurants with lower initial ratings and in dimensions previously flagged as poor by consumers, consistent with the ARMS enhancing restaurant response to negative reviews. We also show that ARMS adoption partially crowds out front-end managerial responses, indicating a substitution between operational remediation and public replies, with the effect strongest when staff take a reflective approach to negative reviews. Finally, adoption is associated with greater consumer engagement, reflected in longer and more detailed reviews. Overall, our results suggest that online review systems, when paired with operational technology, can drive substantive quality improvements beyond mere reputation management.

Guangyu Cao
Peking University
School of Economics
cgy1117@pku.edu.cn

Shenghao He
Peking University
Guanghua School of Management
gsmhsh@pku.edu.cn

Ginger Zhe Jin
Boston University
Questrom School of Business
Markets, Public Policy & Law
and NBER
gzjin@bu.edu

1 Introduction

Consumer reviews posted on digital platforms have become indispensable to both buyers and sellers. By observing others' experiences, consumers learn about the quality of products and services (Fang, 2022; Lewis & Zervas, 2016; Moretti, 2011; Zhao et al., 2013), infer potential risks (Farronato & Zervas, 2026; Mejia et al., 2019), update their beliefs (Acemoglu et al., 2022; Shin et al., 2023; Wu et al., 2015), and ultimately make purchase decisions (Anderson & Magruder, 2012; Chevalier & Mayzlin, 2006; Luca, 2016). For sellers, consumer reviews constitute electronic word-of-mouth (eWOM) and shape their reputation that can substantially affect sales (Anenberg et al., 2022; Basuroy et al., 2003; Chevalier & Mayzlin, 2006; Hollenbeck, 2018; Luca, 2016; Zhu & Zhang, 2010). In practice, the economic importance of reviews has even induced some firms to engage in strategic and sometimes illicit reputation manipulation, such as bribing consumers to write positive reviews, posting fake negative reviews about competitors, or injecting fictitious positive feedback about their own products (Gandhi et al., 2024; He et al., 2022; Lappas et al., 2016; Luca & Zervas, 2016; Mayzlin et al., 2014; Wang et al., 2025).

Beyond the reputation mechanism that primarily influences purchase decisions, consumer reviews, particularly negative ones, contain valuable information about a firm's product and operational deficiencies. If firms can efficiently extract and act upon this information, addressing problems identified by consumers, they can achieve substantive improvements in product and service quality. However, this seemingly intuitive process is far from frictionless. Reviews are posted continuously over time and convey information in multiple formats, including star ratings, images, and textual content that may contain compliments, complaints, sarcasm, or narrative accounts. Dynamic monitoring and extracting actionable insights from such heterogeneous and unstructured information can be challenging. This difficulty is further exacerbated in settings such as the Chinese restaurant industry, which constitutes our empirical sample, where most establishments lack dedicated staff to monitor and analyze consumer reviews. Instead, review monitoring is typically handled sporadically by the outlet manager, often alongside routine operational responsibilities.

This paper documents how a new digital technology can transform consumer reviews into actionable insights for quality improvement. In particular, an automated review monitoring system (ARMS), developed by *Company C*, was introduced to a set of Chinese restaurants in 2019. ARMS enables managers to efficiently capture, interpret, and respond to consumer feedback in real time and at scale. The system continuously scrapes and archives reviews from *Dianping*, the largest review platform in China, and offers two core functions: (i) automated reminders for newly posted negative reviews, and (ii) a work ticket management tool that helps staff systematically address consumer complaints. By leveraging these tools, managers can translate dispersed online feedback into structured operational actions that enhance restaurant quality.

The staggered rollout of ARMS enables a Difference-in-Differences (DID) analysis using detailed consumer reviews and work tickets collected by the system. We employ the two-stage DID estimator proposed by Gardner (2021), which addresses the bias of the two-way fixed-effects OLS (TWFE-OLS) estimator arising from treatment effect heterogeneity in staggered adoption settings. Our estimates show that, following ARMS adoption, the average rating of a restaurant's weekly reviews newly posted on *Dianping* increases by 0.358 stars, compared with a pre-adoption mean of 4.265. The effects are even more pronounced when using alternative measures: The share of negative reviews declines by 0.077 (or 95.1%) from a baseline level of 0.081, while the average sentiment score of review texts increases by 0.119 (or 19.1%) from a baseline of 0.622. Finally, a relative time model confirms that these improvements are not driven by differential pre-adoption trends between treated and control restaurants.

A central concern in interpreting the observed improvements is that ARMS adoption may not have generated *genuine* quality enhancements. Instead, the apparent gains could reflect strategic manipulation of reviews, such as bribing consumers to post positive feedback. To address this concern, we implement three empirical tests. First, if strategic manipulation were the main driver, 5-star reviews would be disproportionately affected. When frontline staff solicit favorable reviews offline, they almost invariably request the highest possible rating to maximize reputational benefits, with little reason to ask for 4- or 4.5-star ratings. Motivated by this, we exclude all 5-star reviews and replicate the analyses. The key results remain stable and significant, indicating that the main findings are not driven by 5-star review inflation. Second, manipulated reviews may be generated from similar templates, which would likely increase textual similarity among 5-star reviews. We find no evidence that textual similarity among 5-star reviews increases following ARMS adoption, a finding that is inconsistent with the manipulation-driven explanation. Third, even if consumers are persuaded to assign a high star rating on *Dianping*, they may still express dissatisfaction in the review text. Such cases, referred to as *masked negativity*, occur when a review appears positive in rating but conveys negative sentiment in its text. If strategic manipulation were predominant, we would expect a significant increase in masked negativity following ARMS adoption. However, the empirical evidence does not support this. Taken together, no evidence suggests that the observed improvements after ARMS adoption are primarily driven by strategic manipulation. In addition, we show that the estimated quality improvements remain robust across a range of alternative specifications, including clustering standard errors at different levels, restricting the analysis to more comparable subsamples, and incorporating controls that proxy for restaurant-level demand fluctuations.

To explore the mechanisms driving the effects of ARMS adoption, we conduct three heterogeneity analyses. First, with respect to initial quality, we find that ARMS fosters convergence in restaurant performance: Establishments with lower pre-adoption ratings experience larger gains, leveraging consumer negative reviews more effectively to improve their overall quality. Second, across different dimensions of the dining experience – including food taste, service, environment, waiting time, price, and food hygiene – we observe that post-adoption improvements are particularly pronounced in areas where restaurants had underperformed prior to adoption. These two patterns align naturally with the design of ARMS as a system for monitoring and acting on negative feedback: Restaurants facing more severe consumer complaints, and dimensions previously flagged by consumers as underperforming, benefit the most from adoption, highlighting the technology's targeted and operationally meaningful impact. Finally, we find that ARMS generates larger improvements among restaurants with shorter histories on *Dianping*, for which current service improvements are more likely to translate into observable changes in platform-displayed star ratings, given their relatively limited stock of historical reviews.

The adoption of ARMS introduces an internal, operationally oriented approach to processing consumer feedback that is not directly observable to the public, which we term *back-end* management. We find that publicly visible managerial responses on the platform, or *front-end* actions, decline following ARMS adoption, indicating that structured back-end workflows can partially substitute for front-end responses. Using discussion texts from work tickets, we further classify back-end staff attitudes toward negative reviews as either *defensive* or *reflective*. Defensive attitudes attribute complaints to consumer-related factors, such as excessive expectations, whereas reflective attitudes emphasize internal service or operational shortcomings. Our results show that ARMS generates significantly smaller improvements in restaurants where staff exhibit more defensive attitudes, highlighting that the effectiveness of technology adoption depends on complementary organizational practices. Moreover, restaurants with defensive back-end attitudes engage more actively in front-end responses and experience a smaller decline in managerial reply rates compared with their reflective counterparts. This pattern suggests that front-end responses capture not only managerial attention to customer feedback but also a reputation-management motive aimed at shaping external perceptions, which may not necessarily correspond to operational follow-ups or actual quality improvements.

Finally, we go beyond restaurant-level outcomes to examine whether ARMS-induced improvements correlate with changes in consumer review behavior. First, we find that consumer reviews became more multidimensional and detailed following ARMS adoption, as evidenced by an increase in the number of dimensions (out of the six) mentioned in each review and a decrease in the proportion of reviews that do not reference any specific dimension. Second, after dining in a restaurant that has adopted ARMS, consumers exerted greater effort when composing and posting reviews on *Dianping*, reflected in more frequent inclusion of pictures and longer texts. Collectively, these patterns indicate that consumers are more engaged in review writing following ARMS adoption, even if they do not observe whether a restaurant has adopted ARMS in the back end. Importantly, prior research has documented that review solicitation or incentivization can degrade the informational quality of consumer reviews (Gao et al., 2025; Khern-am-nuai et al., 2018). Therefore, the observed increase in consumer engagement provides further evidence against a manipulation-driven interpretation.

Our work contributes to several strands of the literature. First, we extend the research on the role of online consumer reviews. While the extant literature has predominantly viewed reviews as consumer-facing signals of quality that influence demand (Acemoglu et al., 2022; Anderson & Magruder, 2012; Chevalier & Mayzlin, 2006; Fang, 2022; Farronato & Zervas, 2026; Lewis & Zervas, 2016; Luca, 2016; Moretti, 2011; Shin et al., 2023; Zhao et al., 2013), we highlight their utility as supply-side feedback for operational improvement. While a subset of the literature has explored seller-side responses, these studies focus primarily on strategic manipulation, specifically, how firms may solicit inauthentic reviews to artificially bolster their reputations (He et al., 2022; Lappas et al., 2016; Luca & Zervas, 2016; Mayzlin et al., 2014; Pocchiari et al., 2024; Wang et al., 2025). Our study diverges from this focus on "gaming" by demonstrating that sellers have strong incentives to address consumer complaints through substantive quality improvements, particularly when digital tools facilitate timely and systematic intervention.

Second, considering the enormous value of consumer reviews, the existing literature has documented three approaches to mitigate the underprovision of consumer reviews: (i) directly motivating consumers by financial incentives and social norms (Burtch et al., 2018; Cabral & Li, 2015; Chen et al., 2010; Li & Xiao, 2014; Yu et al., 2022); (ii) stimulating their desire to write reviews through appropriate platform mechanism design (Deng et al., 2022; Fradkin et al., 2021; Goes et al., 2014; Huang et al., 2019; Jiang et al., 2023); and (iii) guiding consumers to actively engage in review writing through front-end managerial responses (Ananthakrishnan et al., 2023; Chen et al., 2019; Chevalier et al., 2018; Proserpio & Zervas, 2017; Yang et al., 2019). We add to this literature by providing evidence that proactive managerial efforts to make real improvement in the back end can also encourage consumers to engage and provide more informative reviews.

A couple of recent studies convey a similar insight that platform sellers utilize online consumer reviews to improve their products and services (e.g., Ananthakrishnan et al., 2023; Karanam et al., 2025). Our work differs from theirs in two key ways. First, whereas the existing literature primarily studies the front-end handling of consumer reviews, such as whether and how managers make public responses on the platform, we are the first to shed light on back-end management of consumer reviews, especially how consumer reviews are translated into operational follow-up actions to enhance quality improvement, and document the correlation between front- and back-end managerial actions. Second, we highlight the role of digital technology in enabling improvements through unstructured, high-volume online consumer reviews. Intuitively, sellers can learn from consumer reviews, but how they convert rapidly growing textual feedback into concrete operational practices and how this process changes following a major technology adoption remain largely underexplored. This paper seeks to fill this gap.

The remainder of this paper is organized as follows. Section 2 introduces the research background. Section 3 describes the data, variables, and sample construction. Section 4 presents the empirical setting. Section 5 reports the main empirical findings of the impact of ARMS adoption. Section 6 explores the interaction between front-end managerial responses and back-end staff attitudes, as well as ripple effects on consumer review writing. Section 7 concludes and derives managerial implications.

2 Research Background

2.1 Dianping and Consumer Reviews

Many countries host their own dominant consumer review platforms. Yelp plays this role in the United States, while in China the corresponding platform is *Dianping*. Founded in 2003, *Dianping* merged with another major review platform, *Meituan*, in 2015 and has since become the largest and effectively the sole nationwide consumer review platform in China. By 2024, *Dianping* reported coverage of more than 8 million local businesses across virtually all major cities and thousands of smaller cities, and had accumulated 268 million genuine consumer reviews. In China, *Dianping* is deeply embedded in consumers' dining decisions. When planning to eat out, a typical consumer opens the *Dianping* app to search for restaurants, browses ratings and reviews to choose where to dine, and may subsequently leave her own review. As the largest repository of Chinese review texts worldwide, *Dianping* reviews capture consumers' perceptions of restaurants, including complaints, criticism, praise, and expectations, and therefore constitute a valuable source of information for restaurant operators.

Figure 1 displays an illustrative consumer review posted on *Dianping*. To submit a review, a user must provide a star rating on a 5-star scale, while providing a textual comment and attaching photographs are optional. In the sample analyzed in this paper, 97% of individual reviews contain both a star rating and text, while photographs are attached to 79% of individual reviews. Accordingly, the analysis below focuses primarily on variables constructed from star ratings and textual reviews. Importantly, there is no binding constraint linking the star rating to the content of the textual review. As a result, a reviewer may assign the maximum 5-star rating while expressing strong criticism in the accompanying text. This phenomenon, referred to as *masked negativity*, will be employed in the construction of our empirical tests.

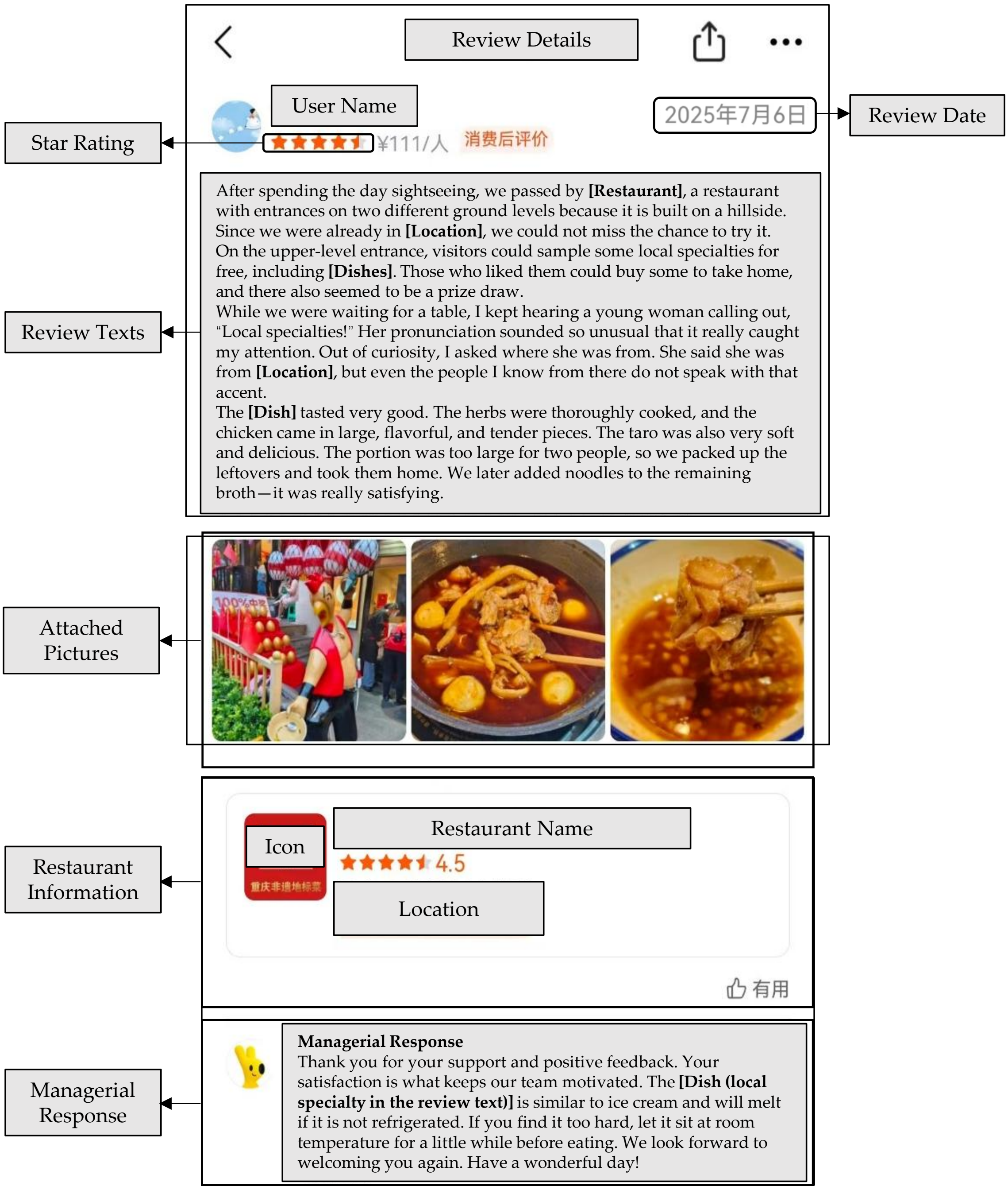


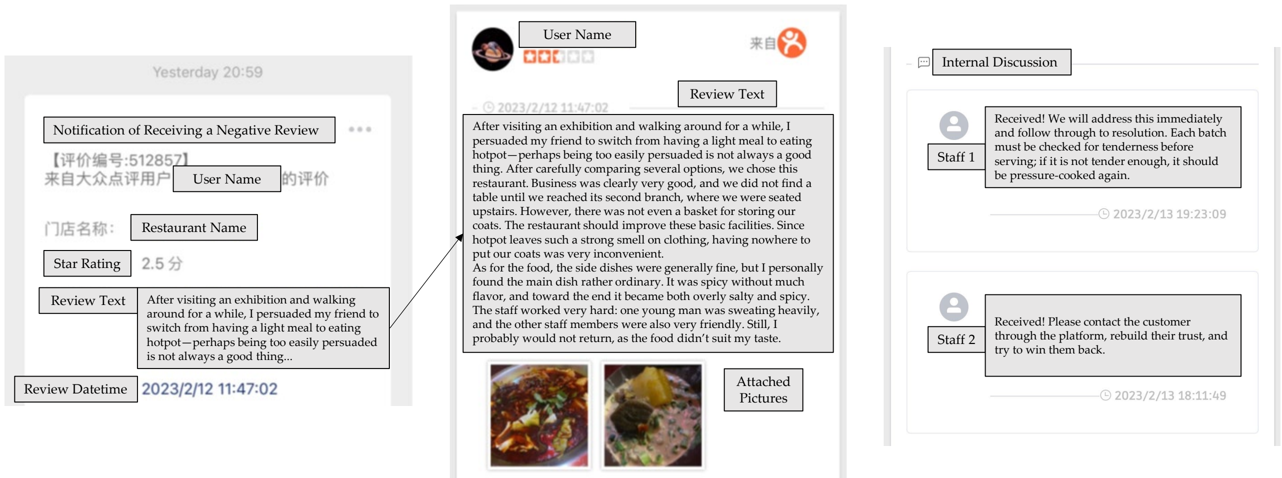
Figure 1. *An Illustrative Example of Consumer Reviews on Dianping*

Notes: This figure presents an illustrative example of consumer reviews on *Dianping*. The original interface and review texts are in Chinese. For readability, we translate the texts into English while preserving the layout. The review writer, reviewed restaurant, locations, and specific dishes are anonymized.

2.2 Automated Review Monitoring System

The Automated Review Monitoring System (ARMS) is a third-party software developed by *Company C*, which operates independently of *Dianping*.[1] ARMS functions only with explicit authorization from business clients, who are typically the headquarters of national or regional restaurant chains. Once authorization is granted, ARMS accesses consumer reviews on *Dianping* for all restaurant outlets owned by the client, storing the retrieved reviews in a private data warehouse. The collected reviews comprise two components. First, ARMS retrospectively retrieves and archives the full set of reviews that existed prior to authorization. Second, from the time of authorization onward, ARMS continuously monitors and captures newly posted reviews, which are scraped and stored on an hourly basis.

Based on the collected reviews, ARMS offers two core functions. First, it generates automated reminders for newly posted negative reviews, identified using both star ratings and textual analysis enabled by machine learning techniques. Second, it provides a work ticket management tool that assists restaurant staff in addressing consumer complaints. To illustrate these functions, Figure 2 presents examples of the ARMS interface as seen by restaurant staff, while Figure 3 depicts the overall workflow. Each function is described in detail below.



(a) (b) (c)

Figure 2. *The ARMS Interface Displayed to Restaurant Staff*

Notes: This figure presents the ARMS interface displayed to restaurant staff. The original interface, review texts and internal discussion texts are in Chinese. For readability, we translate the texts into English while preserving the layout. The review writer and the restaurant are anonymized.

2.2.1 Automated Reminders for Negative Reviews

For each newly posted review, ARMS immediately assesses its content and sends a notification to the restaurant manager and staff whenever either of two conditions is met: (i) The review's star rating

[1]At the company's request, we anonymize its identity and refer to it as *Company C* throughout the paper.

falls below three stars, or (ii) the review text is classified as negative by a textual analysis algorithm developed by *Company C*. This functionality addresses two practical challenges faced by restaurants. First, automatic alerts remove the need for continuous, manual monitoring of reviews on *Dianping*. Second, and more importantly, it enables a systematic analysis of textual content. While star ratings are straightforward to interpret, reading and evaluating review text is time-consuming and prone to subjective judgment. By automating this process, ARMS provides a scalable and consistent way to identify negative feedback based on text, rather than relying solely on star ratings.

Figure 2 Panel (a) illustrates the interface displayed when restaurant staff receive a notification of a negative review. These notifications are pushed directly to staff members' personal mobile phones via WeChat, the dominant messaging application in China. Because WeChat is the primary platform for daily communication and is integrated with multiple functionalities, such as electronic payments and QR code scanning, such notifications are highly salient and difficult to overlook. Upon tapping the notification, restaurant staff are directed to the interface corresponding to the second function, namely work ticket management.

2.2.2 Work Ticket Management

When a staff member taps the notification for a newly detected negative review, he or she is directed to the full review content, as illustrated in Panel (b) of Figure 2. This action automatically generates a corresponding work ticket. As outlined in Figure 3, staff first review the flagged entry to assess whether it contains actionable information. Reviews that lack actionable content, such as a 1-star rating without accompanying text, or that were incorrectly classified as negative by the algorithm may be closed immediately without further action. If, however, the review passes this manual verification, managers can assign specific tasks to designated employees – for example, re-cleaning a table – to address the issue raised by consumers. At the same time, staff can initiate internal discussions directly within the work ticket. Panel (c) of Figure 2 provides a snapshot of these internal processes. Once the issue has been resolved, the manager can close the ticket, completing the feedback loop.

2.2.3 The Process of ARMS Installation

Before discussing the data in detail, it is helpful to clarify the ARMS installation process, as it directly shapes both data availability and sample composition. The collaboration between a firm adopting ARMS and *Company C* involves two key milestones: *authorization* and *adoption*. *Authorization* occurs when a firm grants *Company C* merchant API access on *Dianping*, allowing ARMS to retrieve all historical reviews and continuously capture newly posted reviews for the firm's restaurants. Importantly, authorization alone does not constitute treatment: Providing data access to *Company C* does not alter restaurant operations nor generate any system outputs. By contrast, *adoption* is the treatment exploited

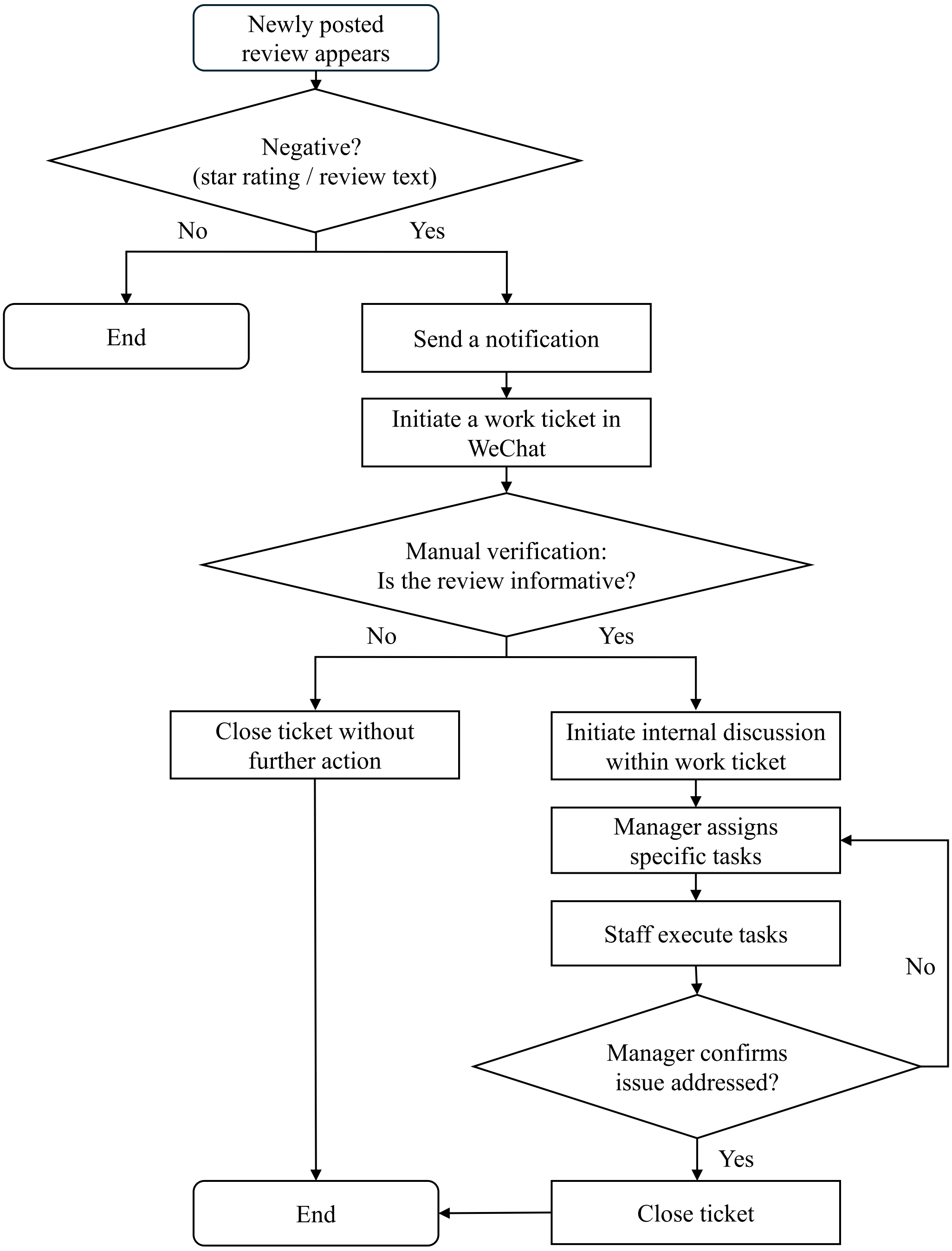


Figure 3. *Illustrative Workflow of ARMS*

in this study. Only when a restaurant adopts ARMS under a subscription plan do the system's functionalities become active. From that point onward, ARMS begins generating notifications for newly posted negative reviews and initiating corresponding work tickets. Reviews posted after authorization but prior to adoption are stored in the private data warehouse as historical records, but do not trigger notifications or work tickets.

Several additional details warrant clarification. First, some firms authorize API access but never adopt ARMS. For these firms, *Company C* continues to retrieve and dynamically capture reviews on a rolling basis, as authorization remains valid throughout the sample period. These never-treated firms form part of the control group in the empirical analysis,[2] for which review data are available, but the treatment – namely, ARMS adoption – is absent. Second, while the automated reminder function is provided by default, the work ticket management tool is optional. Firms may choose to use both functions or only the automated reminders, with no change in subscription fees. Third, ARMS adoption is a top-down decision made at the headquarters level of each firm. The headquarters determines whether to adopt ARMS, and, if so, whether to activate both functions or only the automated alert function. These choices apply to all subordinate restaurant outlets governed by the headquarters. In this sense, ARMS adoption is not driven by the preferences of individual restaurant managers, who must operate within the framework decided by their headquarters.

A natural concern in this context is selection bias: Firms that eventually adopt ARMS may place greater managerial emphasis on consumer reviews, be more responsive to problems highlighted by reviews, or have experienced a higher volume of negative feedback prior to adoption. However, this concern is partially mitigated by the presence of never-treated restaurants. As discussed above, these firms have also authorized *Company C* to access their review data, a step that itself signals the headquarters' attention to and concern for consumer feedback, even though they ultimately chose not to adopt the system. In this respect, never-treated restaurants are likely more comparable to adopting firms than the broader set of firms that have never engaged with ARMS.[3] Moreover, our empirical strategy further addresses potential confounding factors: In addition to restaurant fixed effects, we control for observable characteristics such as location and cuisine type, and allow for time trends or seasonal patterns that may vary across these observables. We have also conducted the analysis using a propensity-score-matched (PSM) sample. Finally, the test of the parallel trends assumption provides

[2]As in a typical staggered DID setting, another component of the control group consists of pre-adoption observations from the treatment group.

[3]In principle, one could scrape *Dianping* reviews for never-treated restaurants with no relationship to *Company C*. We do not pursue this alternative design for two reasons. First, restaurants that have never engaged with *Company C* are likely to place less emphasis on consumer reviews and are therefore less comparable to treated restaurants. Second, without data authorization from restaurants, web scraping cannot recover a restaurant's complete review history. It captures only the reviews visible on *Dianping* at the time of collection, thereby missing reviews that were deleted or revised during our sample period.

an additional check on the comparability of treatment and control groups. We return to a more detailed discussion of potential selection bias when presenting the empirical results.

3 Data and Sample

We compile three main datasets for the empirical analysis in this paper. The first two are provided by *Company C* and include the universe of (i) consumer reviews posted on *Dianping* and (ii) records of work tickets initiated within ARMS. In addition, we collect a set of restaurant-level attributes to account for heterogeneity across establishments. We describe these datasets in detail below.

3.1 Consumer Review

As illustrated in Figure 1 and discussed in Section 2.1, for each consumer review posted on *Dianping*, we extract (i) the star rating, (ii) the full review text, (iii) a dummy variable indicating the presence of attached photographs, and (iv) a dummy variable indicating whether a managerial response is posted. Using both the quantitative and textual information contained in these reviews, we construct a set of variables that fall into three categories: measures of restaurant quality, front-end managerial actions, and consumer review behavior.

Restaurant Quality. The simplest and most direct measure is the star rating itself. Star ratings range from 0.5 to 5 in increments of 0.5. Although imperfect, this measure provides a direct and widely used indicator of consumers' evaluations of restaurant quality. In addition, we construct a dummy variable equal to one if the star rating is less than 3, a threshold commonly used in the restaurant industry to classify "negative reviews." ARMS adopts the same rule, and reviews with ratings less than 3 stars automatically trigger a reminder, as described in Section 2.2.1.

In contrast to the quantitative information conveyed by star ratings, qualitative information contained in review texts typically provides more diagnostic and granular, yet less readily interpretable, signals (Archak et al., 2011; Büschken & Allenby, 2016; Chakraborty et al., 2022; Cho et al., 2022). To quantify this information, we apply a sentiment classification algorithm. The algorithm is a pre-trained model fine-tuned on Chinese restaurant reviews on *Dianping*.[4] The raw outputs of the model are logits, which we transform into probability scores using the softmax function, following standard practice. This procedure yields a sentiment score for each review, taking values between 0 and 1 and representing the probability that the review is positive. This text-based measure captures the overall tone of consumer narratives and provides a richer representation of consumer feedback beyond numeric star ratings.

[4]The model is available at https://huggingface.co/uer/roberta-base-finetuned-dianping-chinese (accessed on September 15, 2025).

In addition to a single measure capturing overall restaurant quality, we further use review texts to construct measures of quality along six distinct dimensions. Based on consultations with experts in the restaurant industry and product managers from *Company C*, we identify six well-established dimensions that capture key aspects of the dining experience: food taste, service, environment, waiting time, price, and food hygiene. For each review, we employ the DeepSeek-V3 large language model to identify whether each dimension is mentioned and, if so, to classify the associated sentiment as positive or negative.[5] This procedure allows us to derive dimension-specific measures of restaurant quality from review texts, complementing the aggregate evaluation.

Front-end Managerial Actions. Restaurant responses to negative reviews can be categorized into two broad types. The first type consists of *back-end actions*, which refer to substantive operational measures undertaken by restaurant staff to improve product or service quality. For example, if customers raise concerns about hygiene, the restaurant may strengthen its cleaning procedures; if customers complain about slow service, the restaurant may accelerate food preparation and delivery. Such operational improvements reflect the original purpose and intended function of ARMS. The second type, by contrast, does not necessarily involve any change in actual operational practices. Instead, restaurant staff may focus solely on managing the review itself, which we term *front-end actions*. For instance, a restaurant may respond to a negative review on *Dianping* with a message such as "We have received your feedback and will strive to improve." In the existing literature, such seller-initiated replies to consumer reviews are commonly referred to as *managerial responses* (Chen et al., 2019; Deng & Ravichandran, 2024; Proserpio & Zervas, 2017). For each consumer review, we construct a dummy variable indicating whether the restaurant posted a response. In the subsequent analysis, this variable serves as our empirical proxy for front-end actions by restaurant staff.

Consumer Review Writing. Finally, because reviews are authored by consumers, we can construct measures that capture changes in consumer review behavior. In the final part of the paper, to examine the ripple effects of ARMS-induced quality improvements on such behavior, we develop a series of variables. The primary aim of these variables is to capture both the richness of review content and the effort consumers exert in writing reviews. To measure content richness, building on the earlier classification into six dimensions, we compute for each review the number of dimensions it addresses; a higher count indicates richer content. Reviews that do not reference any of the six dimensions are classified as *topicless*. To measure consumer effort, we consider two indicators: (i) whether a review includes photographs, which are generally more persuasive and require additional effort, captured by a dummy variable; and (ii) the length of the review text, measured by word count, which reflects the degree of engagement in composing the review.

[5] The prompt for this task is attached in Appendix C.1.

3.2 Work Ticket

The second dataset obtained from ARMS consists of internal discussion texts recorded in work tickets. As illustrated in Figure 3, when a work ticket is triggered by a negative review, restaurant managers and staff may initiate internal discussions to address the issue. These discussions take place within the ARMS interface, and the corresponding communication texts are recorded and stored by the system.

Based on manual coding of the discussion texts and interviews with restaurant staff, we find that, when confronted with negative reviews, staff members' attribution and response strategies can be broadly categorized into two types. Some exhibit a *defensive* attitude, primarily attributing complaints to consumer-related factors, such as excessive expectations or misunderstandings. Others display a *reflective* attitude, focusing on deficiencies in their own service or operational practices.

We again employ the DeepSeek-V3 large language model to classify each discussion into one of these two categories.[6] This classification enables us to examine how staff attitudes toward negative reviews shape the impact of ARMS adoption, even though ARMS itself is a common and neutral managerial tool.

3.3 Restaurant characteristics

The final dataset is self-constructed and comprises a set of cross-sectional variables designed to capture potential heterogeneity across restaurants. Specifically, we use ArcGIS to process raster data on population density and nighttime light intensity and compute their average values within a three-kilometer radius of each restaurant.[7] Both raster datasets are measured in 2018 and therefore predate our sample period, which helps mitigate potential confounding effects arising from ARMS adoption. These two variables are closely related to local demand conditions and economic activity, both of which are important determinants of restaurant operations. In addition, following suggestions from executives at *Company C*, we construct a dummy variable indicating whether a restaurant specializes in hotpot, a highly distinctive and culturally salient dining format in Chinese cuisine. In the subsequent analysis, we collectively refer to these variables as "restaurant attributes" or "restaurant characteristics."

3.4 Sample Construction

Our final sample comprises 122 restaurants. Among them, 104 adopted ARMS during our observation window, while the remaining 18 authorized *Company C* to access their review data but did not adopt ARMS. These two sets of restaurants form the treatment and control groups, respectively, in our empirical framework. Within the treatment group, adoption is not entirely homogeneous. All 104 restaurants activated the automated reminder function, but only 81 additionally implemented the

[6] The prompt for this task is attached in Appendix C.2.

[7] Specifically, the raster data on population density are constructed by *WorldPop*, and the nighttime light intensity data are constructed by Chen et al. (2021).

work ticket management module. As a result, analyses that rely on internal work ticket discussions are conducted on a reduced sample of 99 restaurants, consisting of the 81 fully adopting treatment restaurants and the 18 control restaurants.

Unless otherwise specified, the baseline unit of analysis is restaurant-week. This level of aggregation strikes a balance between noise reduction and temporal resolution: It smooths idiosyncratic daily fluctuations and day-of-week variability, while preserving meaningful short-run variation around adoption. Operationally, we first construct the relevant variables at the individual review level and then aggregate them to restaurant-week by taking averages. Table 1 presents summary statistics for the main variables used in the analysis.

Table 1. *Summary Statistics*

Variable	Obs.	Mean	Std. dev.	Min.	Max.
Panel (a): Restaurant Quality					
Average Star Rating	5,723	4.366	0.568	0.5	5
Proportion of Negative Reviews	5,723	0.065	0.137	0	1
Average Sentiment Score	5,612	0.700	0.200	0.004	0.996
Panel (b): Front-End Actions & Consumer Review Behavior					
Managerial Response Rate	5,723	0.731	0.391	0	1
Average Number of Dimensions Mentioned	5,723	2.286	0.739	0	6
Proportion of Topicless Reviews	5,723	0.101	0.198	0	1
Proportion of Picture-attached Reviews	5,723	0.569	0.338	0	1
Average Length of Reviews (in logs)	5,612	4.396	0.557	0	6.380
Panel (c): Restaurant Characteristics					
Nearby Population Density in 2018	122	15049.77	11519.45	280.43	56138.98
Nearby Nighttime Light Intensity in 2018	122	29.536	10.267	7.169	63.826
Hotpot Dummy	122	0.475	0.501	0	1

Notes: This table reports summary statistics for the main variables. Panel (a) presents measures of restaurant quality, Panel (b) reports variables capturing front-end actions and consumer review behavior, and Panel (c) lists restaurant attributes. The data granularity is restaurant-week in Panels (a) and (b), and restaurant in Panel (c).

ARMS adoption occurred in four concentrated waves: March 2019, February 2023, March 2023, and April 2023. *Company C* granted us access to data spanning 24 weeks before and 24 weeks after each restaurant's adoption date, yielding a symmetric event window of roughly one year centered around adoption. For control restaurants, we observe consumer review data for all calendar weeks that overlap with at least one treatment restaurant's observation window. In other words, the control group's time horizon is defined as the union of the treatment restaurants' observation periods, ensuring temporal comparability between treated and untreated units.

4 Econometric Setting

To evaluate the effects of ARMS adoption on restaurant quality, we exploit its staggered rollout to implement a Difference-in-Differences (DID) design. The baseline specification is as follows:

$$Y_{it} = \alpha_i + \lambda_t + \beta ARMS_{it} + \Gamma(X_i \cdot MoY_t) + \mu G_i \cdot T_t + \varepsilon_{it}, \tag{1}$$

where i and t index restaurants and calendar weeks, respectively. Y_{it} denotes measures of restaurant quality for restaurant i in week t, constructed as described in Section 3.1, including weekly averages of star ratings, the proportion of negative reviews, and sentiment scores. Restaurant and week fixed effects are denoted by α_i and λ_t, respectively. The indicator variable $ARMS_{it}$ equals one if restaurant i has adopted ARMS by week t.

X_i denotes predetermined restaurant characteristics, which are interacted with month-of-year dummies, MoY_t (i.e., January through December). This interaction allows for flexible control of differential time trends across restaurants that may vary systematically with these characteristics. G_i is an indicator for the treatment group restaurant, which interacts with a linear time trend, T_t, to accommodate group-specific trends. ε_{it} denotes the error term. The coefficient of interest, β, captures the effect of ARMS adoption on restaurant quality. For statistical inference, standard errors are clustered at the restaurant level to account for potential serial correlation.

The baseline DID model specified in Equation (1) relies on the parallel trends assumption, namely that, absent ARMS adoption, restaurants in the treatment and control groups would have followed similar outcome trends after we control for observable attributes. To assess the plausibility of this assumption, we estimate a relative time model, in which the adoption indicator $ARMS_{it}$ is decomposed into a set of indicators for time windows relative to the adoption date. Specifically, we estimate the following model:

$$Y_{it} = \alpha_i + \lambda_t + \sum_j \tau_j Pre^j_{it} + \sum_k \omega_k Post^k_{it} + \Gamma(X_i \cdot MoY_t) + \mu G_i \cdot T_t + \varepsilon_{it}. \tag{2}$$

Pre^j_{it} is an indicator variable that equals one if calendar week t is j weeks prior to the ARMS adoption week for restaurant i. Similarly, $Post^k_{it}$ equals one if week t is k weeks after ARMS adoption for restaurant i. All other variables are defined as in Equation (1). Accordingly, the estimated pretreatment coefficients $\{\tau_j\}_{j=-24}^{-1}$ allow us to assess whether treatment and control restaurants exhibited parallel trends prior to ARMS adoption, while the post-adoption coefficients $\{\omega_k\}_{k=0}^{24}$ capture the dynamic effects of ARMS adoption over time.

As emphasized in recent advances in the econometric literature on DID methods, estimating Equations (1) and (2) using TWFE-OLS in a staggered adoption setting may lead to biased estimates for at least two reasons. First, the TWFE-OLS estimator implicitly aggregates multiple two-by-two DID comparisons, potentially assigning negative weights to some comparisons, thereby rendering the overall coefficient difficult to interpret economically (Goodman-Bacon, 2021; Sun & Abraham, 2021). Second, the inclusion of additional controls such as group-specific linear trends (e.g., $G_i \cdot T_t$) can further complicate identification. In TWFE-OLS estimation, these trends are typically estimated using both pre- and post-adoption observations. Consequently, post-adoption treatment effects may mechanically contaminate the estimated counterfactual trends.

To address these concerns, we adopt the two-stage DID estimator proposed by Gardner (2021) as our baseline. This approach resolves the aforementioned issues through a transparent two-step procedure. Intuitively, in the first stage we estimate the model using only untreated observations, thereby identifying fixed effects and trend components without exposure to treatment effects. In the second stage, we residualize the outcome variable for all observations by removing the components estimated in the first stage, and then recover the average treatment effect on the treated (ATT) by regressing the residualized outcomes on the treatment indicator.

5 ARMS Adoption and Restaurant Quality Improvement

In this section, we present empirical evidence on how the adoption of ARMS affects restaurant quality. We begin by reporting results from estimating Equation (1), leveraging both numeric ratings and texts of consumer reviews to construct quality measures. We then turn to the relative time model specification to assess the plausibility of the parallel trends assumption, which is necessary to render a causal interpretation of the DID estimate. Next, we conduct additional empirical tests to address the concern that the observed improvements may be driven by strategic manipulation, such as bribing consumers to post favorable reviews, and implement supplementary analysis to confirm the robustness under alternative specifications. Finally, we conclude this section with a series of heterogeneity analyses to shed light on the potential mechanisms underlying the effects of ARMS adoption.

5.1 ARMS Adoption and Restaurant Quality

We begin by using the most straightforward measure of restaurant quality, the average star rating, as the outcome variable in estimating Equation (1). Table 2 reports the corresponding results, and we progressively introduce additional controls across specifications to demonstrate the robustness of the estimated effects. Column (1) includes only restaurant and week fixed effects. Building on this specification, Column (2) further adds interactions between restaurant attributes and month-of-year dummies, while Column (3) instead incorporates a linear time trend specific to the treatment group.

Finally, Column (4) includes all fixed effects and trend controls as specified in Equation (1).

Table 2. *ARMS Adoption and Quality Improvement*

Dependent Variable:	Average Star Rating			
	(1)	(2)	(3)	(4)
ARMS	0.121*	0.189***	0.269***	0.358***
	(0.062)	(0.070)	(0.087)	(0.095)
Week FE	YES	YES	YES	YES
Restaurant FE	YES	YES	YES	YES
Treatment Trend			YES	YES
Covariates $\times$ MoY		YES		YES
Observations	5723	5723	5723	5723
Clusters	122	122	122	122

Notes: This table reports the results using star rating as the outcome variable. All regressions control for restaurant and week fixed effects. Column (2) further adds interactions between restaurant attributes and month-of-year dummies, while Column (3) instead incorporates a linear time trend specific to the treatment group. Column (4) includes all fixed effects and trend controls. Standard errors are in parentheses and clustered at the restaurant level. *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

Across all specifications, we find a stable and statistically significant improvement in star ratings following the adoption of ARMS. The estimated coefficient β ranges from 0.121 to 0.358, corresponding to a 2.8 – 8.4% increase in weekly average star ratings relative to the pre-adoption sample mean of 4.265. Notably, as we progressively introduce additional controls and trend adjustments, thereby accounting for selection on observables, the magnitude of the estimated coefficient increases. This pattern suggests that potential selection bias likely biases the baseline estimates downward. In other words, restaurants that adopt ARMS may exhibit characteristics associated with relative deterioration in the underlying rating dynamics, so failing to control for such factors would tend to attenuate the estimated treatment effect rather than exaggerate it.

It warrants clarification that this improvement does not translate one-for-one into the star ratings displayed on *Dianping*. Our estimated effect pertains to newly posted reviews, whereas the platform-displayed rating for each restaurant aggregates the full stock of historical reviews. Nonetheless, the results still indicate a systematic upward shift in the flow of incoming consumer evaluations. Over a sufficiently long horizon, sustained improvements in newly posted reviews are likely to accumulate and gradually be reflected in the platform's displayed star ratings. This observation also suggests heterogeneity in restaurants' incentives to improve. For restaurants with a shorter history and a smaller stock of accumulated reviews on *Dianping*, quality improvements are more likely to translate into visible changes in displayed ratings, which in turn may strengthen their incentives to make use of ARMS. We examine this prediction in Section 5.4.3.

The preceding results provide preliminary evidence of a positive impact of ARMS adoption on restaurant quality. However, as emphasized in the existing literature, relying solely on star ratings as outcome measures may be problematic, as they can be subject to rating inflation and strategic distortions (Aziz et al., 2023; Bolton et al., 2019; Filippas et al., 2022; Fradkin et al., 2021; Tadelis, 2016). In light of this concern, we complement the analysis by employing alternative measures of restaurant quality.

First, we compute the proportion of negative reviews at the restaurant-week level, where negative reviews are defined as those with star ratings below 3. This measure more directly captures highly unsatisfactory consumer experiences. As reported in Column (1) of Table 3, the adoption of ARMS is associated with a statistically significant decline in the incidence of negative reviews. The estimated reduction of 7.7 percentage points is a 95.1% decline from the pre-adoption sample mean of 0.081, and thus economically meaningful. This pronounced effect is consistent with the nature of ARMS, which primarily operates through automated alerts triggered by negative reviews. By drawing managerial attention to failures in a timely manner, ARMS is particularly effective at mitigating the occurrence of severe dissatisfaction among consumers.

Table 3. *Alternative Measures of Restaurant Quality*

Dep. Var.	Proportion of Negative Reviews	Average Sentiment Score
	(1)	(2)
ARMS	-0.077***	0.119***
	(0.023)	(0.031)
Week FE	YES	YES
Restaurant FE	YES	YES
Treatment Trend	YES	YES
Covariates × MoY	YES	YES
Observations	5723	5612
Clusters	122	122

Notes: This table shows quality improvements following the adoption of ARMS using alternative measures. Column (1) replaces the dependent variable with the proportion of negative reviews, and Column (2) focuses on the weekly average sentiment score. Standard errors are in parentheses and clustered at the restaurant level. *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

Second, we use review texts to construct a sentiment score as an additional measure of restaurant quality, complementing the previous measures that rely primarily on numerical star ratings. As described in Section 3.1, we employ a sentiment classification algorithm developed by *Dianping* using its proprietary review data, making it particularly well suited to our empirical context, since our primary dataset also consists of reviews obtained from *Dianping*. The sentiment score ranges from 0 to 1 and represents the probability that a given review is positive. Column (2) of Table 3 shows that the weekly average sentiment score increases by 0.119 following ARMS adoption (a 19.1% improvement relative to the pre-treatment baseline of 0.622). This result indicates that, when restaurant quality is assessed using the textual content of consumer reviews rather than numeric ratings alone, the overall tone of consumer feedback becomes systematically more favorable. Taken together, the results in Tables 2 and 3 suggest that ARMS adoption leads to meaningful improvements in the perceived quality of restaurants.

5.2 Tests of Parallel Trends Assumption

To assess the plausibility of the parallel trends assumption, we estimate the relative time model specified in Equation (2) using the two-stage DID estimator. Figure 4 plots the estimated coefficients, along with 95% confidence intervals, using star ratings as the outcome variable. The estimates show no statistically discernible differences in pre-adoption trends between treatment and control restaurants, lending support to the parallel trends assumption. In contrast, following ARMS adoption, positive effects begin to emerge approximately one month (four weeks) after implementation, suggesting that the system requires time to generate measurable improvements. Over a longer horizon, these effects persist, indicating a sustained impact of ARMS. We also estimate the relative time model using the proportion of negative reviews and the sentiment score as outcome variables; the corresponding results, presented in Figures A1 and A2, exhibit highly similar pre-adoption patterns and post-adoption dynamics.

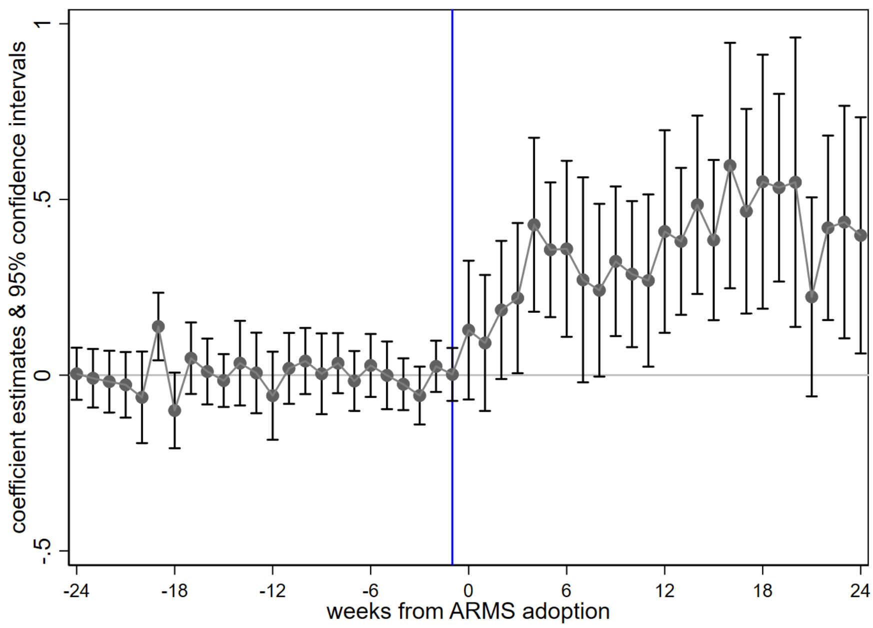


Figure 4. *Dynamic Effects of ARMS Adoption on Star Ratings*

Notes: This figure presents estimates from the relative time model specified in Equation (2) using the two-stage DID estimator. The *x*-axis denotes the number of weeks relative to ARMS adoption. Points indicate coefficient estimates, and vertical bars represent 95% confidence intervals.

5.3 Robustness Checks

So far, we have shown that, following the adoption of ARMS, restaurants experienced significant improvements across multiple dimensions, including higher star ratings, a lower proportion of negative reviews, and more favorable text-based sentiment scores. Collectively, these findings suggest that ARMS enables restaurant staff to more effectively process information contained in negative reviews and translate consumer feedback into tangible improvements in product and service quality. We now proceed with robustness checks along two dimensions. The first addresses the alternative explanation that the observed improvements are driven by strategic manipulation, while the second examines a series of alternative econometric specifications.

5.3.1 Addressing Concerns about Strategic Manipulation

A key threat to the above interpretation, or an alternative explanation for the observed patterns, is strategic manipulation. Specifically, the installation of ARMS may not have led to real quality improvement; instead, the apparent improvement could reflect manipulative practices, such as bribing consumers to post positive reviews or otherwise distorting review content (He et al., 2022; Lappas et al., 2016; Luca & Zervas, 2016; Mayzlin et al., 2014; Wang et al., 2025). If this were the case, the implications of our empirical findings would be fundamentally different. Below, we present a series of empirical tests that, while not fully conclusive, provide evidence against the hypothesis that the observed improvements are primarily driven by strategic manipulation.

Excluding 5-Star Reviews. 5-star reviews warrant particular attention because, if strategic manipulation is indeed present, it should disproportionately manifest in this rating category. The intuition is straightforward: When frontline staff solicit favorable reviews from customers through offline persuasion, they almost invariably request the highest possible rating in order to maximize the reputational benefits; there is little incentive for staff to ask for a 4- or 4.5-star rating, as such ratings do not fully serve the purpose of inflating visible reputation metrics.

Motivated by this consideration, we conduct a test to mitigate the potential influence of strategic manipulation by excluding all 5-star reviews from the sample. We then recompute the average star rating, the proportion of negative reviews, and the text-based sentiment score, and replicate the main analyses using these adjusted measures. If the estimated effects of ARMS adoption persist after removing 5-star reviews, this would suggest that the observed quality improvements are unlikely to be driven primarily by strategic manipulation. Consistent with our expectation, Table 4 presents the corresponding results, which show that our findings remain robust even when all 5-star reviews are excluded.

Table 4. Analysis Excluding 5-Star Reviews

Dep. Var.	Average Star Rating	Proportion of Negative Reviews	Average Sentiment Score
	(1)	(2)	(3)
ARMS	0.451***	-0.116***	0.143***
	(0.132)	(0.040)	(0.044)
Week FE	YES	YES	YES
Restaurant FE	YES	YES	YES
Treatment Trend	YES	YES	YES
Covariates × MoY	YES	YES	YES
Observations	5348	5348	5264
Clusters	122	122	122

Notes: This table reports results from analyses that exclude 5-star reviews. For some restaurant-week observations, all reviews are 5 stars, leading to the exclusion of these observations from the regression. This explains the slightly smaller sample size compared with Tables 2 and 3. Standard errors are in parentheses and clustered at the restaurant level. *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

Textual Similarity among 5-Star Reviews. The second test is motivated by another common practice used by restaurants to influence consumer reviews. When encouraging customers to post favorable feedback, staff may provide suggested wording or templates to reduce the effort required to write a review. As a result, manipulated 5-star reviews are likely to exhibit greater textual similarity to one another (Gu et al., 2025). Accordingly, if the observed improvements in star ratings and review sentiment were primarily driven by such manipulation, we would expect textual similarity among 5-star reviews to increase following ARMS adoption.

Considering this possibility, we construct a test based on 5-star reviews with non-empty text. Specifically, we represent each review using TF-IDF character n-gram vectors and then construct three restaurant-week measures of textual similarity.[8] The first measure is the average pairwise cosine similarity, which compares every pair of 5-star reviews posted for the same restaurant in the same week. A higher value indicates that 5-star reviews within a restaurant-week tend to use more similar language. The second measure is the proportion of duplicated reviews. Specifically, we identify review pairs with cosine similarity above 0.90 and calculate their share among all 5-star review pairs within the same restaurant-week. The third measure is the average similarity to the "centroid" review. For each restaurant-week, we define the centroid review as the average TF-IDF vector across all 5-star reviews, which captures their common wording pattern, and then compute the similarity between each review and this centroid. The results reported in Table 5 show no significant increase in any of these measures following ARMS adoption. This pattern provides suggestive evidence against the concern that our

[8]Following common practice in textual similarity analysis, we first clean the review text by removing spaces, line breaks, tabs, and other irrelevant characters. Review texts are represented using character-level TF-IDF vectors that incorporate sequences of two to four consecutive Chinese characters.

results are driven by an influx of coordinated manipulated reviews.

Table 5. *ARMS Adoption and Textual Similarity among 5-Star Reviews*

Dep. Var.:	Textual Similarity among 5-Star Reviews		
	Pairwise	Duplicated	Centroid
	(1)	(2)	(3)
ARMS	0.001	0.000	-0.029
	(0.013)	(0.013)	(0.042)
Week FE	YES	YES	YES
Restaurant FE	YES	YES	YES
Treatment Trend	YES	YES	YES
Covariates $\times$ MoY	YES	YES	YES
Observations	4107	4107	4107
Clusters	122	122	122

Notes: This table examines the potential impact of ARMS adoption on textual similarity among 5-star reviews. The similarity measures are constructed at the restaurant-week level using all 5-star reviews with non-empty text. We represent each review text using TF-IDF character n-gram vectors and compute within-restaurant-week textual similarity among 5-star reviews with texts. The columns report three alternative measures: (i) the average pairwise cosine similarity across all review pairs, (ii) the proportion of duplicated (cosine similarity $\geq$ 0.90) review pairs, and (iii) the average similarity between each review and the restaurant-week centroid vector. For some restaurant-week observations, there are at most one 5-star review with texts, leading to the exclusion of these observations from the regression. This explains the smaller sample size compared with Tables 2 and 3. Standard errors are in parentheses and clustered at the restaurant level. *, **, and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

Masked Negativity in Consumer Reviews. Compared with manipulating star ratings, altering review text is substantially more costly, so the textual sentiment score provides a more credible reflection of consumers' authentic perceptions, which we have already employed as an outcome variable. This feature also offers a way to test for the potential influence of strategic manipulation. Even if consumers are persuaded to give a high star rating on *Dianping* and actually agree to do so to avoid offending restaurant staff, they may still express dissatisfaction in the review text. When this occurs, it results in *masked negativity*, defined as cases in which a review appears positive in star rating but conveys negative sentiment in the text. If the observed improvements were primarily driven by strategic manipulation, we would expect to see a substantial increase in masked negativity after ARMS adoption.

Concretely, we define *masked negativity* as reviews that exhibit a star rating above a specified threshold but a sentiment score below a given threshold. For star ratings, we consider both 5-star (corresponding to a "perfect experience" by *Dianping*'s definition) and 4-star (corresponding to a "good experience") as thresholds. For sentiment scores, we experiment with 0.3 and 0.7 as cutoffs. Under each combination of thresholds (i.e., star rating = 5 & sentiment score < 0.3, star rating $\geq$ 4 & sentiment score < 0.3, star rating = 5 & sentiment score < 0.7, star rating $\geq$ 4 & sentiment score < 0.7), we identify reviews meeting the criteria for masked negativity and compute their proportion at the restaurant-week level.

Using this measure, we estimate the impact of ARMS adoption on the proportion of masked negativity. As reported in Table 6, we do not find any significant increase in masked negativity following ARMS adoption. The only exception is Column (2), where the coefficient is significantly negative rather than positive, which contradicts the hypothesis that strategic manipulation drives our previous findings. Overall, these results provide no evidence that ARMS has induced manipulation of consumer reviews in the form of masked negativity.

Table 6. *ARMS Adoption and Masked Negativity*

Dep. Var.	Proportion of Masked Negativity in Positive Reviews			
	(1)	(2)	(3)	(4)
ARMS	0.010	-0.048**	0.019	-0.035
	(0.022)	(0.022)	(0.048)	(0.049)
Rating Threshold	5	4	5	4
Sentiment Threshold	0.30	0.30	0.70	0.70
Week FE	YES	YES	YES	YES
Restaurant FE	YES	YES	YES	YES
Treatment Trend	YES	YES	YES	YES
Covariates $\times$ MoY	YES	YES	YES	YES
Observations	5165	5619	5165	5619
Clusters	122	122	122	122

Notes: This table examines the potential impact of ARMS adoption on masked negativity in consumer reviews. Each column corresponds to a different combination of thresholds used to define positive star ratings and negative sentiment in review texts. Standard errors are in parentheses and clustered at the restaurant level. *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

5.3.2 Alternative Econometric Settings

We have also conducted a series of robustness checks using alternative model specifications, with the results reported in Table B1. First, to account for potential intra-firm correlation among restaurants belonging to the same firm, we adjust the clustering level of the standard errors from the restaurant to the firm level. Given that the number of firms is naturally smaller than the number of subordinate restaurants, we employ a wild cluster bootstrap method (Cameron et al., 2008; Webb, 2023) to avoid biased standard errors and unreliable inference arising from a limited number of clusters. As shown in Column (1), the estimated effect remains statistically significant after switching the clustering unit.

Second, although the relative time model supports the general comparability of treatment and control restaurants, concerns may remain about residual differences between the two groups. To address this issue, we first restrict the sample to treatment restaurants only and conduct a before-and-after comparison within this group. As shown in Column (2), after abstracting from potential differences between treatment and control restaurants, the estimated effect remains stable in both magnitude and statistical significance. We further complement this exercise with a propensity score matching (PSM) analysis. Specifically, we estimate each restaurant's propensity to adopt ARMS using pre-treatment restaurant characteristics and construct a 1:1 matched sample. Column (3) reports the corresponding regression results. The estimated effect remains similar in both sign and magnitude, suggesting that our findings are unlikely to be driven by observable imbalances between treatment and control restaurants. Additional details on the PSM procedure, along with results under alternative matching specifications, are reported in Appendix B.2.

Third, another potentially omitted factor is demand shocks faced by restaurants, which could influence the consumer reviews they receive. To partially address this concern, we additionally control for the one-period lagged review volume, which serves as a proxy for demand-side conditions. As shown in Column (4), although the estimated effect slightly decreases in magnitude, the impact of ARMS adoption on restaurant quality remains positive and statistically significant.

5.4 Heterogeneity Analyses

5.4.1 Convergence in Restaurant Quality

At the time of ARMS adoption, restaurants may be operating at markedly different levels of business performance. For instance, some may be heavily burdened by negative reviews, whereas others may already perform relatively well in terms of consumer feedback. Given that ARMS primarily facilitates improvement by directing restaurant staff's attention to negative reviews, a natural hypothesis is that restaurants facing more severe negative feedback stand to benefit more from ARMS adoption.

To empirically examine this hypothesis, we ask *Company C* to compute each treatment restaurant's pre-adoption star rating, calculated from all historical consumer reviews prior to ARMS adoption. We then standardize this measure by demeaning and dividing by its standard deviation, yielding a cross-sectional index denoted by $PreRating_i$. Using the same specification as Column (4) of Table 2, we further interact the treatment indicator $ARMS_{it}$ with $PreRating_i$. The coefficient on this interaction term captures heterogeneity in the effects of ARMS adoption as a function of restaurants' initial performance.

The regression results are reported in Table 7, where the outcome variables are, respectively, the star rating, the proportion of negative reviews, and the sentiment score. For restaurants starting from a relatively low level of initial quality (i.e., $PreRating_i = 0$), the improvements following ARMS adoption are both statistically significant and economically meaningful, as reflected in the coefficients on $ARMS_{it}$. By contrast, as indicated by the estimated coefficients on the interaction term across the three columns, restaurants with higher pre-adoption ratings benefit significantly less from ARMS adoption. The results indicate that ARMS promotes convergence in restaurant quality by enabling initially disadvantaged restaurants to achieve greater improvements through effective use of consumer negative reviews.

Table 7. *Convergence in Restaurant Quality after ARMS Adoption*

Dep. Var.	Average Star Rating	Proportion of Negative Reviews	Average Sentiment Score
	(1)	(2)	(3)
ARMS	0.358***	-0.077***	0.119***
	(0.093)	(0.023)	(0.031)
ARMS × PreRating	-0.214***	0.043***	-0.028*
	(0.037)	(0.009)	(0.015)
Week FE	YES	YES	YES
Restaurant FE	YES	YES	YES
Treatment Trend	YES	YES	YES
Covariates × MoY	YES	YES	YES
Observations	5723	5723	5612
Clusters	122	122	122

Notes: This table illustrates convergence in restaurant quality following ARMS adoption. In addition to the original explanatory variable, we include its interaction with a measure of the restaurant's initial quality. Standard errors are in parentheses and clustered at the restaurant level. *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

5.4.2 Dimensional Improvements

As introduced in Section 3.1, beyond the measure capturing overall restaurant quality, we identify whether each of the six well-established dimensions is mentioned in a review and, if so, whether the associated sentiment is positive or negative: food taste, service, environment, waiting time, price, and food hygiene. This more granular analysis of consumer review data allows us to examine heterogeneity along another dimension, namely, which aspects of restaurant operations benefit most from ARMS adoption.

Specifically, for each dimension k, we estimate the following equation:

$$Quality_{it}^{k} = \alpha_i + \lambda_t + \beta^k ARMS_{it} + \eta^k ARMS_{it} \times PreNegProp_i^k + \Gamma(X_i \cdot MoY_t) + \mu G_i \cdot T_t + \epsilon_{it}, \quad (3)$$

where the dependent variable $Quality_{it}^{k}$ is the dimensional quality measure, defined as the proportion of reviews containing positive comments on dimension k for restaurant i in week t. The variable $PreNegProp_i^k$ denotes the proportion of reviews containing negative comments on dimension k received by restaurant i prior to ARMS adoption, constructed in the same spirit as $PreRating_i$.

Table 8 reports the results, with each column corresponding to one dimension. Across all dimensions, we find positive effects of ARMS adoption, manifested as increases in the proportion of reviews expressing positive feedback along the corresponding dimension. Moreover, the coefficients on the interaction terms indicate that these improvements are dimension-specific. In particular, for dimensions in which restaurants performed relatively poorly prior to adoption (i.e., those associated with higher $PreNegProp_i^k$), the post-adoption improvements are more pronounced.

Table 8. *Dimensional Improvements*

Dep. Var.	Proportion of Positive Reviews on Dimension k					
	Environment	Service	Food Taste	Waiting Time	Price	Food Hygiene
	(1)	(2)	(3)	(4)	(5)	(6)
ARMS	0.138*	0.293***	0.097*	0.234***	0.205**	0.121**
	(0.078)	(0.067)	(0.052)	(0.059)	(0.089)	(0.062)
ARMS $\times PreNegProp_i^k$	0.027	0.082***	0.014	0.063***	0.111***	0.123***
	(0.034)	(0.025)	(0.022)	(0.023)	(0.025)	(0.023)
Week FE	YES	YES	YES	YES	YES	YES
Restaurant FE	YES	YES	YES	YES	YES	YES
Treatment Trend	YES	YES	YES	YES	YES	YES
Covariates $\times$ MoY	YES	YES	YES	YES	YES	YES
Observations	5723	5723	5723	5723	5723	5723
Clusters	122	122	122	122	122	122

Notes: This table depicts dimensional improvements by examining whether ARMS brings about a larger impact on dimensions where the restaurant used to perform poorly. The dependent variable is weekly dimensional quality measured by the proportion of reviews containing positive feedback on that dimension. Standard errors are in parentheses and clustered at the restaurant level. *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

This pattern is weaker for the environment and food taste dimensions, as indicated by interaction estimates that are statistically indistinguishable from zero, although their signs remain positive. This is plausible, as improvements in the dining environment often require substantial fixed investments and may therefore be less flexible in the short run, while food taste is a fundamental attribute of a restaurant's offering and may leave limited scope for adjustment. For example, a restaurant specializing in spicy cuisine may find it difficult to respond to requests for lighter dishes.

5.4.3 *Dianping* History and Quality Improvement

As noted earlier when discussing the baseline results, improvements reflected in newly posted reviews may carry different implications for restaurants with different histories on *Dianping*. For restaurants with a longer platform history and a larger accumulated stock of reviews, displayed ratings are likely to be less sensitive to current improvements in service quality. By contrast, restaurants with a shorter history may face higher marginal returns to acting on customer feedback, as service improvements can be reflected more quickly in their observed ratings and review outcomes. This reasoning suggests that ARMS may generate larger improvements for restaurants with shorter *Dianping* histories.

To examine this possibility, we construct an indicator variable, $LongHistory_i$, for treatment restaurants, equal to one if a restaurant's *Dianping* history exceeds the sample mean and zero otherwise. We then augment the baseline specification in Column (4) of Table 2 by interacting the treatment indicator with $LongHistory_i$. Table 9 reports the results. The estimated coefficients on the interaction term in Columns (1) and (2) indicate that restaurants with longer *Dianping* histories experience significantly smaller improvements following ARMS adoption, although the coefficient loses statistical significance in the last column. Overall, these findings are consistent with the idea that restaurants with shorter *Dianping* histories face higher marginal returns to acting on customer feedback and are therefore more motivated to leverage ARMS.

Table 9. *Dianping History and Quality Improvement*

Dep. Var.	Average Star Rating	Proportion of Negative Reviews	Average Sentiment Score
	(1)	(2)	(3)
ARMS	0.410***	-0.090***	0.117***
	(0.102)	(0.024)	(0.033)
ARMS × LongHistory	-0.153*	0.037**	0.007
	(0.081)	(0.019)	(0.027)
Week FE	YES	YES	YES
Restaurant FE	YES	YES	YES
Treatment trend	YES	YES	YES
Covariates × MoY	YES	YES	YES
Observations	5723	5723	5612
Clusters	122	122	122

Notes: This table shows heterogeneous effect of ARMS-induced quality improvement by *Dianping* history. In addition to the original explanatory variable, we include its interaction with an indicator that the restaurants *Dianping* history exceeds the sample mean. Standard errors are in parentheses and clustered at the restaurant level. *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

6 Managerial Actions and Consumer Review Behavior

6.1 Managerial Actions: Front-End and Back-End

Research on online reputation management has largely focused on actions directly observable to consumers, such as how managers publicly respond to consumer reviews on digital platforms (Ananthakrishnan et al., 2023; Chen et al., 2019; Chevalier et al., 2018; Deng & Ravichandran, 2024; Proserpio & Zervas, 2017; Yang et al., 2019). Yet ARMS introduces a complementary, less visible channel through which restaurants process negative feedback: internal discussion, coordination, and operational improvements. We refer to these internally oriented efforts as *back-end* managerial actions, in contrast to *front-end* responses posted publicly on the platform. In practice, managers may allocate effort across both channels, which are not mutually exclusive, but higher back-end engagement may reduce resources available for front-end responses due to typical constraints in manpower and time. By examining both front-end and back-end actions, this section explores how internal organizational factors, such as staff attitudes toward negative feedback, shape the effectiveness of technology adoption.

6.1.1 ARMS Adoption and Front-End Actions

We begin by analyzing front-end actions, as exemplified by managerial responses, the practice whereby restaurant staff may choose to reply to consumer reviews on the platform, making these replies visible to all users browsing *Dianping*. Specifically, we calculate the weekly proportion of reviews that receive a response from restaurants on *Dianping* and then evaluate the impact of ARMS adoption. Table 10 reports the results, organized according to the same logic as Table 2. Consistent with the baseline findings, we observe across all specifications that ARMS partially crowds out public-facing replies, suggesting that the structured back-end workflow introduced by ARMS can, to some extent, substitute for front-end review management. [9]

Table 10. *ARMS Impact on Managerial Response Rate*

Dependent Variable:	Managerial Response Rate			
	(1)	(2)	(3)	(4)
ARMS	-0.226***	-0.283***	-0.381***	-0.475***
	(0.044)	(0.093)	(0.047)	(0.072)
Week FE	YES	YES	YES	YES
Restaurant FE	YES	YES	YES	YES
Treatment Trend			YES	YES
Covariates $\times$ MoY		YES		YES
Observations	5723	5723	5723	5723
Clusters	122	122	122	122

Notes: This table reports changes in the front-end managerial response rate following ARMS adoption. The specifications in each column correspond to those in the matching columns of Table 2. Standard errors are in parentheses and clustered at the restaurant level. *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

6.1.2 Back-End Staff Attitudes and Quality Improvement

The work ticket data, which record restaurant staff discussions regarding issues raised in negative reviews, provide a unique opportunity to unpack the back-end processes through which a restaurant improves its operations. Specifically, our empirical analysis focuses on staff attitudes toward consumer negative reviews. As detailed in Section 3.2, based on the discussion texts in work tickets, these attitudes can be broadly classified as *defensive* or *reflective*. The former reflects a tendency to attribute complaints to consumer-related factors, such as excessive expectations, whereas the latter emphasizes

[9] However, we do not find a significant change in the types of reviews managers respond to or the information contained in their responses.

internal service or operational shortcomings. These differentiated attitudes are likely to influence the extent to which restaurants can translate ARMS adoption into tangible quality improvements.

To empirically examine this, we compile all discussion texts at the restaurant level and classify each text as either *defensive* or *reflective* using the DeepSeek-V3 large language model. For each restaurant, we then calculate the proportion of texts exhibiting a defensive attitude and construct an indicator variable, $Defensive_i$, equal to one if this proportion exceeds the sample average. We interact $Defensive_i$ with the treatment dummy and re-estimate the DID model. Two caveats are worth noting. First, because only 81 restaurants activated the work ticket management tool, the analysis is conducted on a reduced sample. Second, work ticket data are available only after ARMS adoption, meaning that $Defensive_i$ is based on post-treatment information. Including such a variable on the right-hand side of the regression raises a classic "bad control" concern (Angrist & Pischke, 2009). Consequently, while these estimates are informative about the role of back-end staff attitudes in shaping ARMS effectiveness, they should be interpreted with caution and do not warrant a straightforward causal interpretation.

The results are reported in Columns (1) – (3) of Table 11. We find that ARMS generates significantly smaller improvements for restaurants where internal discussions of negative reviews exhibit a more defensive orientation. Quantitatively, the quality gains for restaurants with a defensive attitude are less than half of those observed for their more reflective counterparts, whether measured by star ratings or by the proportion of negative reviews. These findings indicate that the returns to ARMS are not mechanical consequences of system installation or activation. Instead, they are related to whether frontline staff use the system as an input for shortcoming recognition and continuous improvement, rather than responding by deflecting blame. From a managerial perspective, an important accompanying task when deploying tools such as ARMS is therefore to pair technological rollout with organizational complements that shape how employees interpret and act upon the information surfaced by the technology. In short, effective technology adoption requires integration with organizational change, rather than standalone implementation.

Building on the preceding results on front-end managerial responses and back-end staff attitudes, a natural follow-up question concerns the relationship between the two. To provide suggestive evidence, we again use the managerial response rate as the outcome variable and add an interaction term between the ARMS adoption dummy and $Defensive_i$. Column (4) of Table 11 reveals an intriguing pattern: Although managerial response rates decline after adoption for both groups, the decline is slightly smaller for the defensive group than for the reflective group. In other words, restaurants whose staff exhibit more reflective, self-oriented attitudes toward negative reviews experience a more pronounced reduction in front-end managerial responses following ARMS adoption.

Table 11. Back-End Staff Attitudes and Quality Improvement

Dep. Var.	Average Star Rating	Proportion of Negative Reviews	Average Sentiment Score	Managerial Response Rate
	(1)	(2)	(3)	(4)
ARMS	0.510***	-0.109***	0.161***	-0.510***
	(0.136)	(0.031)	(0.042)	(0.085)
ARMS × Defensive	-0.279**	0.061**	-0.053	0.131*
	(0.116)	(0.026)	(0.035)	(0.074)
Week FE	YES	YES	YES	YES
Restaurant FE	YES	YES	YES	YES
Treatment Trend	YES	YES	YES	YES
Covariates × MoY	YES	YES	YES	YES
Observations	4840	4840	4743	4840
Clusters	99	99	99	99

Notes: This table examines how staff attitudes influence the quality improvement associated with ARMS adoption. Restaurants are classified as *defensive* if the proportion of defensive discussion texts in work tickets exceeds the sample average. The regression sample is restricted to treatment restaurants with available internal discussion records and to control restaurants. Standard errors are in parentheses and clustered at the restaurant level. *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

This observed negative correlation between reflective back-end attitudes and front-end managerial responses suggests that the allocation of managerial effort in the wake of digital technologies like ARMS is shaped by internal factors such as organizational culture. When staff are more open to self-attribution and learning from operational shortcomings, a technology that facilitates the systematic use of negative reviews can shift attention away from purely front-end actions, such as posting public responses, toward back-end remediation. This insight also underscores that, in the literature on managerial responses, changes in observed reply rates should be interpreted with caution. Managerial responses are not a straightforward measure of attention or learning; rather, as our results indicate, back-end attitudes can shape the effort managers devote to front-end actions, complicating the interpretation of empirical findings in this domain.

6.2 Ripple Effects on Consumer Engagement in Review Writing

Toward the end of this paper, we go beyond restaurant-level outcomes to investigate whether ARMS-induced improvements are reflected in changes in consumer review behavior. Because ARMS adoption is a back-end decision that is unobservable to consumers, any differences in review behavior between ARMS-adopting and non-adopting restaurants are likely to capture the effects of actual consumer experiences, which may in turn be influenced by ARMS-driven operational improvements. Specifically, we focus on two dimensions of review behavior: the informativeness of reviews and the effort consumers invest in composing them.

We first examine the informativeness of consumer reviews, for which we construct two measures. The first measure is the average number of dimensions (out of six) mentioned in a review. The second is the proportion of reviews that do not reference any specific dimension, which are largely uninformative in terms of content. Columns (1) and (2) of Table 12 present regression results using these two measures as outcome variables. Together, the estimates indicate that, following ARMS adoption, consumer reviews became more multidimensional and detailed, reflecting an increase in both the richness and informational value of review content.

Table 12. *Ripple Effects on Consumer Engagement in Review Writing*

Dep. Var.	Average Dimension Count	Topicless Proportion	Picture-attached Proportion	Log Average Length
	(1)	(2)	(3)	(4)
ARMS	0.235*	-0.049	0.172***	0.290***
	(0.138)	(0.035)	(0.046)	(0.104)
Week FE	YES	YES	YES	YES
Restaurant FE	YES	YES	YES	YES
Treatment Trend	YES	YES	YES	YES
Covariates $\times$ MoY	YES	YES	YES	YES
Observations	5723	5723	5723	5612
Clusters	122	122	122	122

Notes: This table reports the ripple effects of ARMS on consumer engagement in review writing. The dependent variable in Column (1) is the number of dimensions mentioned in review texts, and Column (2) replaces it with the proportion of reviews that do not mention any dimension. In Column (3), the dependent variable is the proportion of reviews with attached pictures. Column (4) examines the weekly average log length of review texts. Standard errors are in parentheses and clustered at the restaurant level. *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

Another aspect of consumer review behavior is the effort exerted in writing reviews, which we measure by (i) whether consumers attach pictures to their reviews, and (ii) the length of the review text. Correspondingly, Columns (3) and (4) report results using the proportion of reviews with attached pictures and the logarithm of review length as outcome variables, respectively. The estimates show positive effects on both dimensions, indicating that consumers exerted greater effort when composing and posting reviews on *Dianping*.

We acknowledge that informativeness and consumer effort are closely intertwined. Writing more informative reviews generally requires greater effort, and, conversely, longer and more effortful reviews are typically more informative. Therefore, the results in Table 12 should be interpreted as evidence of changes in consumer review behavior, observed from complementary angles, rather than as definitive interpretations of any single variable.

Additionally, the observed impact on consumer review writing provides further evidence that our findings are unlikely to be driven purely by strategic manipulation. Prior research has documented that review solicitation or incentivization can degrade the information quality of consumer reviews (Gao et al., 2025; Khern-am-nuai et al., 2018). Consequently, if the post-ARMS improvements were primarily the result of strategic manipulation, such as bribing consumers to leave 5-star ratings, we would expect consumers to expend less effort and be more perfunctory in writing reviews. In contrast, as shown in Table 12, consumers appear to be more engaged in drafting reviews, which directly counters a manipulation-driven interpretation.

7 Conclusive Discussion and Managerial Implications

This paper studies how the automated review monitoring system transforms online feedback into operational improvements. Exploiting the staggered adoption of ARMS, we find that adoption increases average ratings, reduces the proportion of negative reviews, and improves textual sentiment. These gains are not driven by review manipulation: Results are stable when excluding 5-star reviews, and we do not observe an increase in textual similarity among 5-star reviews or masked negativity. We further document heterogeneity: ARMS effects are stronger for restaurants with lower initial quality and shorter *Dianping* history, and improvements are larger in dimensions where restaurants historically performed poorly.

Additional analysis provides implications on managerial actions and consumer engagement in review writing. First, as a tool to strengthen back-end management, the adoption of ARMS substitutes for front-end managerial response. Second, the benefits are substantially muted when back-end attitudes exhibit a more defensive style. Third, back-end attitudes are systematically related to front-end response behavior. Finally, consumers write more detailed and multidimensional reviews following the adoption of ARMS, which indicates higher engagement and suggests that reviews are less likely to be subject to manipulation-driven distortions.

Our findings yield several managerial implications for restaurants and, more broadly, for firms seeking to leverage user-generated content for operational improvement. First, the role of ARMS highlights that the central challenge in the digital age is not the lack of consumer information, but the ability to make publicly available information actionable. Online review platforms such as *Dianping* already generate vast volumes of consumer feedback, yet such information is often dispersed, noisy, and difficult to process in real time. ARMS adds value not by creating new information, but by transforming existing public information into timely, structured, and decision-relevant inputs for managers. This suggests that, under conditions of information overload, managerial bottlenecks increasingly lie in attention, prioritization, and execution rather than in information acquisition per se. For practitioners, investing in systems that reduce cognitive and organizational frictions in processing feedback may yield returns comparable to, or even exceeding, investments in traditional quality monitoring technologies.

Second, our evidence on back-end staff attitudes underscores that technology alone is insufficient to generate performance gains. Using work ticket data, we show that restaurants whose internal discussions of negative reviews are more reflective rather than defensive experience significantly larger quality improvements following ARMS adoption. This finding implies that the effectiveness of digital monitoring tools critically depends on how frontline employees and managers interpret and respond to the information these tools surface. Defensive reactions that attribute complaints to consumer bias or unreasonable expectations can blunt the learning potential of consumer feedback, even when such feedback is delivered in a timely and organized manner. From a managerial perspective, this highlights the importance of complementary organizational practices, such as training, incentive design, and leadership communication, that foster a learning-oriented mindset. Technology adoption should therefore be paired with deliberate efforts to shape employee attitudes toward feedback, rather than treated as a purely technical upgrade.

Third, the observed ripple effects on consumer review writing suggest the existence of a virtuous cycle between operational improvement and consumer engagement. We find that after ARMS adoption, consumer reviews become more informative, more detailed, and more effortful. This pattern is consistent with a feedback loop in which genuine improvements in service quality motivate consumers to invest more effort in articulating their experiences, which in turn generates richer information for firms to learn from and further improve. For managers, this implies that quality improvement initiatives may generate indirect benefits by enhancing the informational value of subsequent consumer feedback, thereby amplifying long-run returns. For platforms such as *Dianping*, these results point to a novel mechanism to stimulate high-quality user participation: by facilitating merchants' effective utilization of consumer feedback, platforms may indirectly encourage consumers to contribute more substantive and informative reviews, improving the overall quality of platform content.

Taken together, our findings suggest that the managerial value of digital review systems lies not only in aggregating information like providing star ratings but in enabling learning, shaping organizational responses, and fostering positive feedback loops between firms and consumers. Effective deployment of such systems requires aligning technology like ARMS, organizational culture, and platform design to fully realize their potential.

References

Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2022). Learning from Reviews: The Selection Effect and the Speed of Learning. *Econometrica*, *90*(6), 2857–2899.

Ananthakrishnan, U., Proserpio, D., & Sharma, S. (2023). I Hear You: Does Quality Improve with Customer Voice? *Marketing Science*, *42*(6), 1143–1161.

Anderson, M., & Magruder, J. (2012). Learning from the Crowd: Regression Discontinuity Estimates of the Effects of An Online Review Database. *Economic Journal*, *122*(563), 957–989.

Anenberg, E., Kuang, C., & Kung, E. (2022). Social Learning and Local Consumption Amenities: Evidence from Yelp. *Journal of Industrial Economics*, *70*(2), 294–322.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science*, *57*(8), 1485–1509.

Aziz, A., Li, H., & Telang, R. (2023). The Consequences of Rating Inflation on Platforms: Evidence from a Quasi-Experiment. *Information Systems Research*, *34*(2), 590–608.

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How Critical Are Critical Reviews? The Box Office Effects of Film Critics, Star Power, and Budgets. *Journal of Marketing*, *67*(4), 103–117.

Bolton, G. E., Kusterer, D. J., & Mans, J. (2019). Inflated Reputations: Uncertainty, Leniency, and Moral Wiggle Room in Trader Feedback Systems. *Management Science*, *65*(11), 5371–5391.

Burtch, G., Hong, Y., Bapna, R., & Griskevicius, V. (2018). Stimulating Online Reviews by Combining Financial Incentives and Social Norms. *Management Science*, *64*(5), 2065–2082.

Büschken, J., & Allenby, G. M. (2016). Sentence-Based Text Analysis for Customer Reviews. *Marketing Science*, *35*(6), 953–975.

Cabral, L., & Li, L. (2015). A Dollar for Your Thoughts: Feedback-Conditional Rebates on eBay. *Management Science*, *61*(9), 2052–2063.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics*, *90*(3), 414–427.

Chakraborty, I., Kim, M., & Sudhir, K. (2022). Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure and Missing Attributes. *Journal of Marketing Research*, *59*(3), 600–622.

Chen, W., Gu, B., Ye, Q., & Zhu, K. X. (2019). Measuring and Managing the Externality of Managerial Responses to Online Customer Reviews. *Information Systems Research*, *30*(1), 81–96.

Chen, Y., Harper, F. M., Konstan, J., & Li, S. X. (2010). Social Comparisons and Contributions to Online Communities: A Field Experiment on Movielens. *American Economic Review*, *100*(4), 1358–1398.

Chen, Z., Yu, B., Yang, C., Zhou, Y., Yao, S., Qian, X., Wang, C., Wu, B., & Wu, J. (2021). An Extended Time Series (2000–2018) of Global NPP-VIIRS-like Nighttime Light Data from a Cross-Sensor Calibration. *Earth System Science Data*, *13*(3), 889–906.

Chevalier, J. A., Dover, Y., & Mayzlin, D. (2018). Channels of Impact: User Reviews when Quality is Dynamic and Managers Respond. *Marketing Science*, *37*(5), 688–709.

Chevalier, J. A., & Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research*, *43*(3), 345–354.

Cho, H. S., Sosa, M. E., & Hasija, S. (2022). Reading between the Stars: Understanding the Effects of Online Customer Reviews on Product Demand. *Manufacturing & Service Operations Management*, *24*(4), 1977–1996.

Deng, C., & Ravichandran, T. (2024). Managerial Response to Online Positive Reviews: Helpful or Harmful? *Information Systems Research*, *35*(4), 1802–1823.

Deng, Y., Zheng, J., Khern-Am-Nuai, W., & Kannan, K. (2022). More than the Quantity: The Value of Editorial Reviews for a User-Generated Content Platform. *Management Science*, *68*(9), 6865–6888.

Fang, L. (2022). The Effects of Online Review Platforms on Restaurant Revenue, Consumer Learning, and Welfare. *Management Science*, *68*(11), 8116–8143.

Farronato, C., & Zervas, G. (2026). Consumer Reviews and Regulation: Evidence from New York City Restaurants. *Marketing Science*, *45*(4), 752–772.

Filippas, A., Horton, J. J., & Golden, J. M. (2022). Reputation Inflation. *Marketing Science*, *41*(4), 733–745.

Fradkin, A., Grewal, E., & Holtz, D. (2021). Reciprocity and Unveiling in Two-Sided Reputation Systems: Evidence from an Experiment on Airbnb. *Marketing Science*, *40*(6), 1013–1029.

Gandhi, A., Hollenbeck, B., & Li, Z. (2024). Misinformation and Mistrust: The Equilibrium Effects of Fake Reviews on Amazon.com. *Working Paper*.

Gao, B., Wang, J., Ding, X., & Guo, Y. (2025). The Pitfalls of Review Solicitation: Evidence from a Natural Experiment on TripAdvisor. *Management Science*, *71*(2), 1671–1691.

Gardner, J. (2021). Two-Stage Differences in Differences. *Working Paper*.

Goes, P. B., Lin, M., & Au Yeung, C.-m. (2014). "Popularity Effect" in User-Generated Content: Evidence from Online Product Reviews. *Information Systems Research*, *25*(2), 222–238.

Goodman-Bacon, A. (2021). Difference-in-Differences with Variation in Treatment Timing. *Journal of Econometrics*, *225*(2), 254–277.

Gu, X., Cao, J., & Fang, Y. (2025). Review Manipulation and Filtering on Digital Platforms. *Information Systems Research*, *37*(2), 927–947.

He, S., Hollenbeck, B., & Proserpio, D. (2022). The Market for Fake Reviews. *Marketing Science*, *41*(5), 896–921.

Hollenbeck, B. (2018). Online Reputation Mechanisms and the Decreasing Value of Chain Affiliation. *Journal of Marketing Research*, *55*(5), 636–654.

Huang, N., Burtch, G., Gu, B., Hong, Y., Liang, C., Wang, K., Fu, D., & Yang, B. (2019). Motivating User-Generated Content with Performance Feedback: Evidence from Randomized Field Experiments. *Management Science*, *65*(1), 327–345.

Jiang, L. D., Ravichandran, T., & Kuruzovich, J. (2023). Review Moderation Transparency and Online Reviews: Evidence from A Natural Experiment. *MIS Quarterly*, *47*(4), 1693–1708.

Karanam, S. A., Agarwal, A., & Barua, A. (2025). Follow Your Heart or Listen to Users? The Case of Mobile App Design. *Information Systems Research*, *36*(3), 1846–1870.

Khern-am-nuai, W., Kannan, K., & Ghasemkhani, H. (2018). Extrinsic Versus Intrinsic Rewards for Contributing Reviews in an Online Platform. *Information Systems Research*, *29*(4), 871–892.

Lappas, T., Sabnis, G., & Valkanas, G. (2016). The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry. *Information Systems Research*, *27*(4), 940–961.

Lewis, G., & Zervas, G. (2016). The Welfare Impact of Consumer Reviews: A Case Study of the Hotel Industry. *Working Paper*.

Li, L., & Xiao, E. (2014). Money Talks: Rebate Mechanisms in Reputation System Design. *Management Science*, *60*(8), 2054–2072.

Luca, M. (2016). Reviews, Reputation, and Revenue: The Case of Yelp.com. *Working Paper*.

Luca, M., & Zervas, G. (2016). Fake it Till You Make it: Reputation, Competition, and Yelp Review Fraud. *Management Science*, *62*(12), 3412–3427.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional Reviews: An Empirical Investigation of Online Review Manipulation. *American Economic Review*, *104*(8), 2421–2455.

Mejia, J., Mankad, S., & Gopal, A. (2019). A for Effort? Using the Crowd to Identify Moral Hazard in New York City Restaurant Hygiene Inspections. *Information Systems Research*, *30*(4), 1363–1386.

Moretti, E. (2011). Social Learning and Peer Effects in Consumption: Evidence from Movie Sales. *Review of Economic Studies*, *78*(1), 356–393.

Pocchiari, M., Schoenmueller, V., & Dover, Y. (2024). Online Review Updating: Prevalence and Implications for Platforms and Businesses. *Working Paper*.

Proserpio, D., & Zervas, G. (2017). Online Reputation Management: Estimating the Impact of Management Responses on Consumer Reviews. *Marketing Science*, *36*(5), 645–665.

Shin, D., Vaccari, S., & Zeevi, A. (2023). Dynamic Pricing with Online Reviews. *Management Science*, *69*(2), 824–845.

Sun, L., & Abraham, S. (2021). Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. *Journal of Econometrics*, *225*(2), 175–199.

Tadelis, S. (2016). Reputation and Feedback Systems in Online Platform Markets. *Annual Review of Economics*, *8*(1), 321–340.

Wang, L., Luo, X., Qiu, L., Xu, F., & Cui, X. (2025). Win by Hook or Crook? Self-Injecting Favorable Online Reviews to Fight Adjacent Rivals. *Information Systems Research*, *36*(3), 1333–1353.

Webb, M. D. (2023). Reworking Wild Bootstrap-Based Inference for Clustered Errors. *Canadian Journal of Economics*, *56*(3), 839–858.

Wu, C., Che, H., Chan, T. Y., & Lu, X. (2015). The Economic Value of Online Reviews. *Marketing Science*, *34*(5), 739–754.

Yang, M., Zheng, Z., & Mookerjee, V. (2019). Prescribing Response Strategies to Manage Customer Opinions: A Stochastic Differential Equation Approach. *Information Systems Research*, *30*(2), 351–374.

Yu, Y., Khern-am-nuai, W., & Pinsonneault, A. (2022). When Paying for Reviews Pays Off: The Case of Performance-Contingent Monetary Rewards. *MIS Quarterly*, *46*(1), 609–626.

Zhao, Y., Yang, S., Narayan, V., & Zhao, Y. (2013). Modeling Consumer Learning from Online Product Reviews. *Marketing Science*, *32*(1), 153–169.

Zhu, F., & 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.

A Tests of Parallel Trends Using Alternative Outcomes

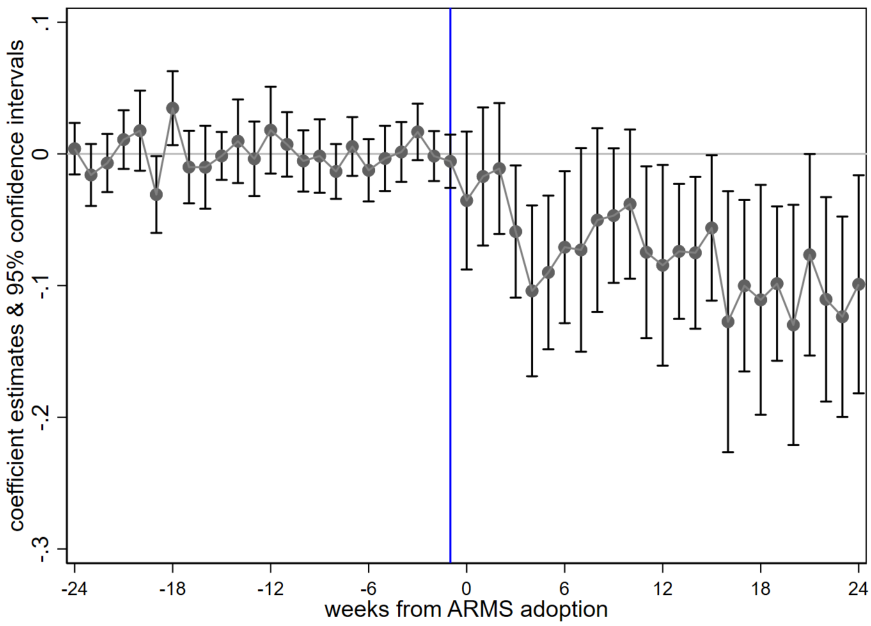


Figure A1. *Dynamic Effects of ARMS Adoption on the Proportion of Negative Reviews*

Notes: This figure presents estimates from the relative time model specified in Equation (2) using the two-stage DID estimator. The x-axis denotes the number of weeks relative to ARMS adoption. Points indicate coefficient estimates, and vertical bars represent 95% confidence intervals.

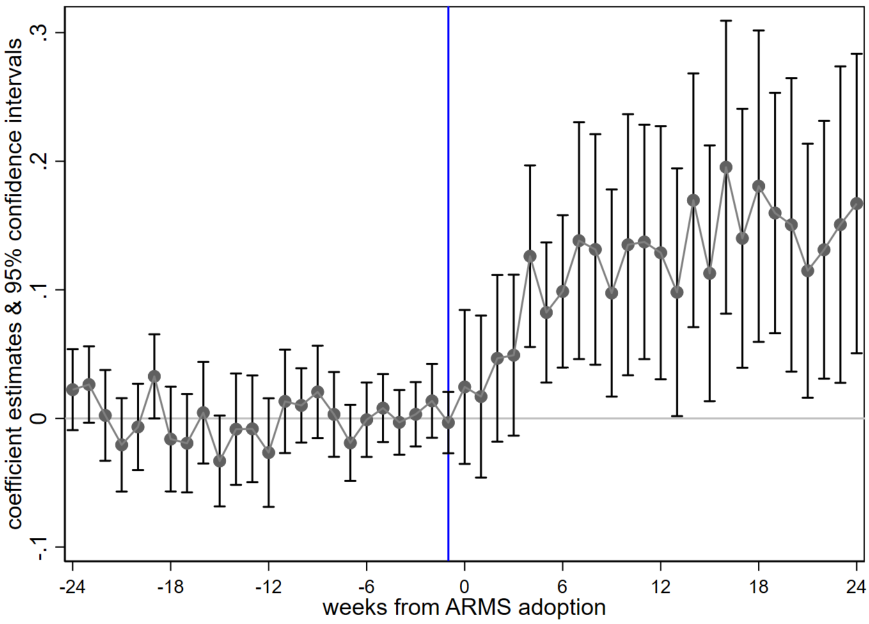


Figure A2. *Dynamic Effects of ARMS Adoption on the Sentiment Score*

Notes: This figure presents estimates from the relative time model specified in Equation (2) using the two-stage DID estimator. The x-axis denotes the number of weeks relative to ARMS adoption. Points indicate coefficient estimates, and vertical bars represent 95% confidence intervals.

B Robustness Checks

B.1 Results under Alternative Econometric Settings

Table B1. *Robustness Check Using Alternative Model Specifications*

Dep. Var.	Average Rating Score			
	(1)	(2)	(3)	(4)
ARMS	0.358**	0.451**	0.285***	0.194**
	[0.021]	(0.214)	(0.089)	(0.083)
Week FE	YES	YES	YES	YES
Restaurant FE	YES	YES	YES	YES
Treatment Trend	YES		YES	YES
Covariates $\times$ MoY	YES	YES	YES	YES
Only Treatment Restaurants		YES		
Propensity Score Matching			YES	
Lag # of Reviews				YES
Observations	5723	4361	4415	5723
Clusters	10	104	89	122

Notes: This table reports results from alternative specifications to assess robustness. Column (1) replicates the baseline specification with standard errors clustered at the firm level. In Column (2), the linear time trend specific to treatment group restaurants is omitted, as only the treatment group is used in the regression. Column (3) reports the estimates with 1:1 matched sample. Column (4) further includes a one-period lag of the number of consumer reviews to proxy for demand-side shocks faced by restaurants. The p-value from the wild cluster bootstrap method (Cameron et al., 2008; Webb, 2023) is reported in brackets in Column (1). Standard errors in parentheses are clustered at the restaurant level for Columns (2), (3) and (4). *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

B.2 Details of Propensity Score Matching

B.2.1 Matching Procedure

This section presents additional evidence to address the concern on the comparability of treatment and control restaurants. We conduct propensity score matching before re-estimating the DID specification. Specifically, we estimate a restaurant-level logit model in which the covariates include restaurant characteristics mentioned in Section 3.3 and the baseline mean of the outcome variable. The fitted values from this model are used as propensity scores. We then construct matched samples using five alternative matching procedures: 1:1, 1:2, 1:3 nearest-neighbor matching, and kernel matching, all with replacement. Because the number of control restaurants is limited, the matching procedures allow control restaurants to be used as matches for multiple treated restaurants. We therefore incorporate the resulting matching weights in the subsequent DID regressions. After constructing each matched sample, we return to the restaurant-week panel and re-estimate our baseline DID specification on the matched sample.

B.2.2 Covariate Balance

To assess whether propensity score matching improves the comparability between treatment and control restaurants, we examine covariate balance before and after matching. Table B2 reports the balance checks for the covariates used in the propensity score model. For each matching procedure, means after matching are evaluated using the corresponding matching weights. The matched samples generally exhibit smaller differences than the unmatched sample. This pattern suggests that the matching procedures improve balance in observable restaurant characteristics.

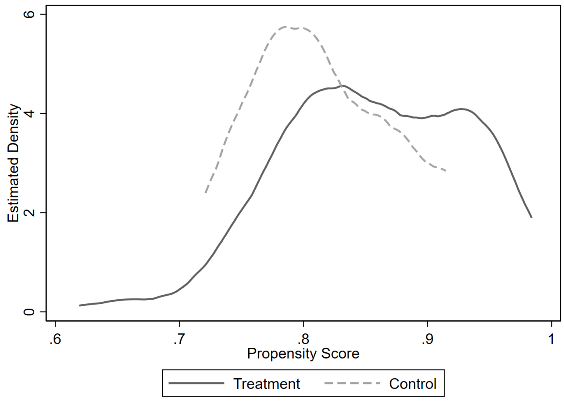
B.2.3 Common support

Figure B1 examines the common support of the estimated propensity scores before and after matching. Panel (a) plots the propensity score distributions for treatment and control restaurants before matching. Panels (b) – (e) show the corresponding propensity score distributions after applying alternative matching procedures. Relative to the unmatched sample, the matched samples display improved overlap between treatment and control restaurants. This suggests that the matching procedures improve comparability between treated restaurants and observationally similar control restaurants.

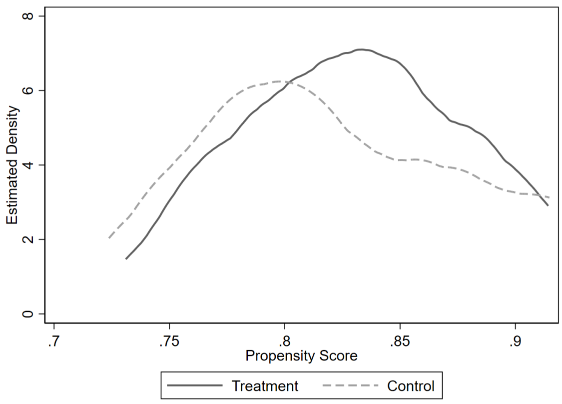
B.2.4 Matched-sample two-stage DID estimates

Table B3 reports the two-stage DID estimates using matched samples. Regressions preserve the same specification as the baseline analysis in Equation (1) and incorporate matching weights. Across all matching procedures, the estimated coefficient on ARMS remains positive and statistically significant. These results indicate that the estimated improvement in restaurant quality is not driven by observable differences between treatment and control restaurants before ARMS adoption.

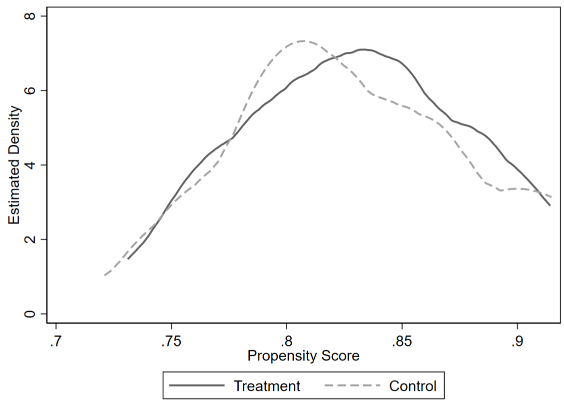
As an additional diagnostic, Figure B2 provides tests of parallel trends assumption by estimating the relative time model specified in Equation (2) with matching weights. Across all matching procedures, the estimates provide no evidence of differential pre-adoption trends between treatment and control restaurants, supporting the validity of the DID framework in the matched samples. Moreover, the estimated effects persist throughout the 24-week post-adoption horizon, further confirming the robustness of our findings.



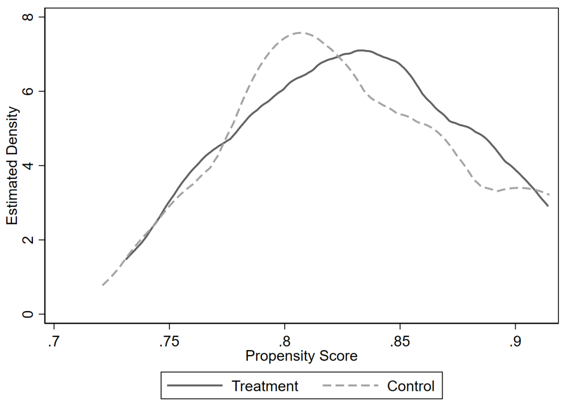
(a) Before Matching



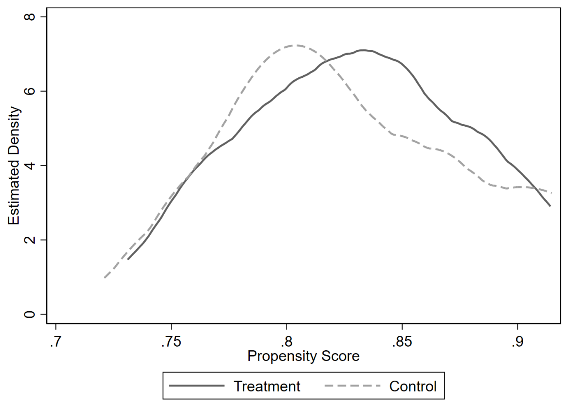
(b) 1:1 NN Matching



(c) 1:2 NN Matching



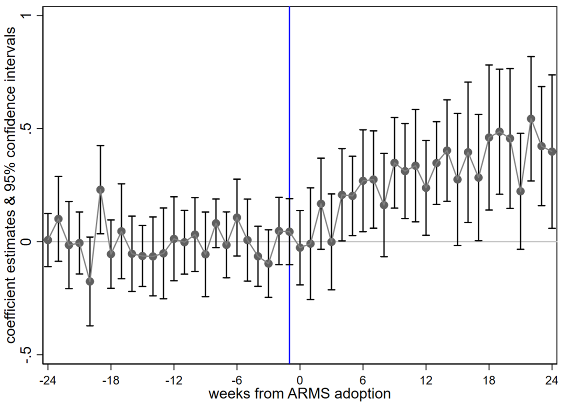
(d) 1:3 NN Matching



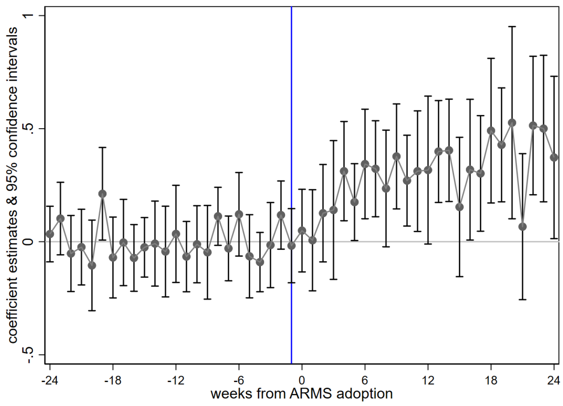
(e) Kernel Matching

Figure B1. *Common Support of Estimated Propensity Scores Before and After Matching*

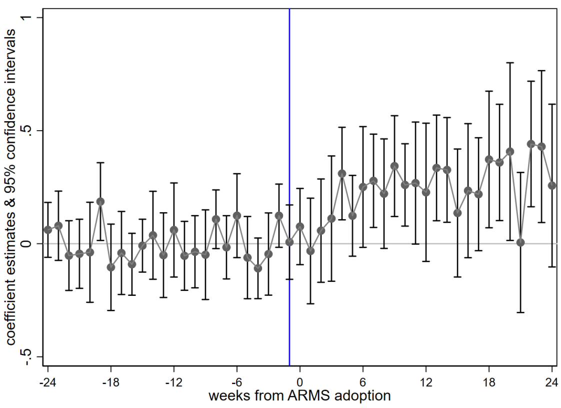
Notes: This figure plots kernel density estimates of the estimated propensity scores for treatment and control restaurants before and after propensity score matching. Panel (a) presents the unmatched sample. Panels (b) – (e) present the matched samples obtained from 1:1, 1:2, 1:3 nearest-neighbor matching, and kernel matching, respectively. The propensity score is estimated from a restaurant-level logit model using restaurant characteristics and the baseline mean of weekly star rating. For the matched samples, the density estimates incorporate the corresponding matching weights.



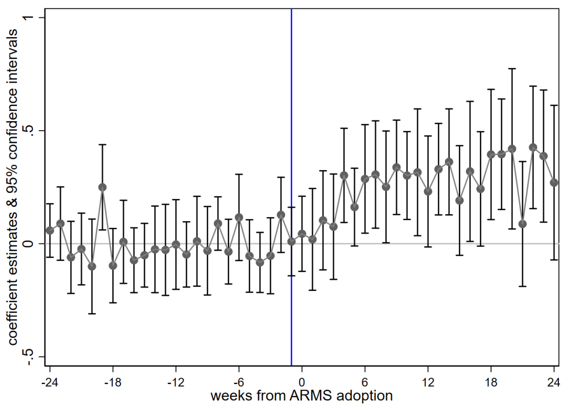
(a) *1:1 NN Matching*



(b) *1:2 NN Matching*



(c) *1:3 NN Matching*



(d) *Kernel Matching*

Figure B2. *Dynamic Effects with Propensity Score Matching*

Notes: This figure presents estimates from the relative time model specified in Equation (2) with matching weights using the two-stage DID estimator. Panels (a) – (d) present the estimates obtained from 1:1, 1:2, 1:3 nearest-neighbor matching, and kernel matching, respectively. The x-axis denotes the number of weeks relative to ARMS adoption. Points indicate coefficient estimates, and vertical bars represent 95% confidence intervals.

Table B2. *Comparison of Covariates Before and After Matching*

Covariate	Treatment (mean)	Treatment (N)	Control (mean)	Control (N)	p-value of t test for difference
Panel (a): Before Matching					
Nearby population density	15345	104	13344	18	0.499
Nearby nighttime light intensity	30.198	104	25.705	18	0.086
Hotpot dummy	0.510	104	0.278	18	0.070
Baseline mean weekly star rating	4.274	104	4.307	18	0.616
Panel (b): 1:1 Nearest-neighbor Matching					
Nearby population density	13065	72	11706	17	0.297
Nearby nighttime light intensity	25.802	72	26.130	17	0.673
Hotpot dummy	0.333	72	0.264	17	0.366
Baseline mean weekly star rating	4.284	72	4.298	17	0.716
Panel (c): 1:2 Nearest-neighbor Matching					
Nearby population density	13065	72	11999	18	0.401
Nearby nighttime light intensity	25.802	72	26.003	18	0.795
Hotpot dummy	0.333	72	0.264	18	0.366
Baseline mean weekly star rating	4.284	72	4.277	18	0.844
Panel (d): 1:3 Nearest-neighbor Matching					
Nearby population density	13065	72	12162	18	0.477
Nearby nighttime light intensity	25.802	72	26.310	18	0.506
Hotpot dummy	0.333	72	0.264	18	0.366
Baseline mean weekly star rating	4.284	72	4.292	18	0.830
Panel (e): Kernel Matching					
Nearby population density	13065	72	12068	18	0.441
Nearby nighttime light intensity	25.802	72	25.880	18	0.921
Hotpot dummy	0.333	72	0.280	18	0.489
Baseline mean weekly star rating	4.284	72	4.280	18	0.907

Notes: This table reports covariate balance before and after propensity score matching under different matching procedures. Means after matching are evaluated using the corresponding matching weights.

Table B3. *DID Estimates with Propensity Score Matching*

Dependent Variable:	Average Star Rating		
	(1)	(2)	(3)
ARMS	0.297***	0.240***	0.262***
	(0.085)	(0.085)	(0.085)
Week FE	YES	YES	YES
Restaurant FE	YES	YES	YES
Treatment trend	YES	YES	YES
Covariates $\times$ MoY	YES	YES	YES
Matching procedure	1:2 NN	1:3 NN	Kernel
Observations	4469	4469	4469
Clusters	90	90	90

Notes: This table reports the DID estimates with propensity score matching using star rating as the outcome variable. Columns (1) – (3) report results based on 1:2 nearest-neighbor matching, 1:3 nearest-neighbor matching, and kernel matching, respectively. Matching weights are assigned at the restaurant level and applied to all restaurant-week observations. Standard errors are in parentheses and clustered at the restaurant level. *, ** and *** indicate statistical significance at 10%, 5% and 1% level, respectively.

C Prompts for the Large Language Model

C.1 Dimensions and Corresponding Sentiment

The prompt (English translated version) we provided to Deepseek-V3 for classifying dimensions mentioned and corresponding sentiment tendencies is:

> You are a professional review-text dimension classification bot. Please process the text according to the rules below:
>
> **Analysis Rules**
>
> 1. Check whether the text mentions any of the following 6 dimensions: ["Dining environment", "Service", "Food taste", "Waiting time" , "Price", "Food hygiene"]
>
> 2. Sentiment judgment (ONLY for mentioned dimensions):
>
> Positive: contains explicit positive words/phrases (e.g., "comfortable", "friendly", "delicious", "good value", etc.)
>
> Negative: contains explicit negative words/phrases (e.g., "dirty/messy", "indifferent", "tastes bad", "overcharging", etc.)
>
> Mixed (both positive and negative): the same dimension includes both positive and negative descriptions (e.g., "The staff were friendly, but they miscalculated the bill.").
>
> **Output Specification**
>
> The output should be a list of three values.
>
> First value: whether this dimension is mentioned (0 for not mentioned and 1 for mentioned)
>
> Second value: whether this dimension contains any positive statements (0 for no and 1 for yes)
>
> Third value: whether this dimension contains any negative statements (0 for no and 1 for yes)
>
> **An example**
>
> *Review text*: "The environment in the store is quite nice, and there are many people coming to eat here. There are free desserts, and the [**dish name**] is particularly delicious. I strongly recommend it. There are small oranges and some snacks. The taste is good, but the business is too good, and the dishes are served a bit slowly. The service is still very enthusiastic, and the seasonings are complete. Overall, it's quite comfortable to eat here."
>
> *Result*: environment: [1,1,0], service: [1,1,0], food_taste: [1,1,0], waiting_time: [1,0,1], price: [0,0,0], food_hygiene: [0,0,0]

C.2 Back-End Staff Attitudes

The prompt (English translated version) we provided to Deepseek-V3 for classifying back-end staff attitudes based on internal discussion texts is:

You are a professional text classification bot. Your task is to classify a summarized negative-review incident description for a restaurant. For each input text, generate a binary variable: defensive, following the rules below.

Classification Rules

Set defensive = 1 if the text indicates that the restaurant attributes the incident to the customer or the customer's behavior, and the proposed solution focuses on addressing the customer directly (e.g., contacting the customer to request deletion of the review, compensating or appeasing the customer, persuasion, negotiation, or other customer-facing interventions). Otherwise, if the text indicates that: the restaurant attributes the incident to its own issues (e.g., service failure, operational mistakes, food quality problems, hygiene problems, management issues), and the proposed solution focuses on improving the restaurant's own performance, set defensive = 0.

An Example

Discussion Text: "We promptly processed the customer's use of the voucher. The customer took a long time to scan the QR code, place the order and select the dishes, while the dishes for the customers at the adjacent table who placed their orders quickly were served much faster."

Result: defensive = 1.