

A Blessing or a Curse? The Impact of Platform-initiated Comment Moderation on Subsequent Answer Generation on Social Media Platform

Ran (Alan) Zhang
Texas Tech University
ran.alan.zhang@ttu.edu

Yuanhong Ma
Beihang University
yuanhongma@buaa.edu.cn

Xingyue (Luna) Zhang
University of Washington Tacoma
xyzhang5@uw.edu

Yong Tan
University of Washington
ytan@u.washington.edu

Abstract

Social media moderation encompasses actions undertaken by platforms to uphold community norms. While existing literature predominantly examines the direct impact of moderation on reducing harmful behavior, limited attention has been given to its spillover effect on other content, particularly on unmoderated content. This study exploits a temporary shutdown of the commenting function on a large Q&A platform to investigate the spillover effect of comment moderation on the subsequent answer generation. Our analysis reveals that comment moderation induce a decrease in the volume of the subsequent answers, with an improved quality during and after the shutdown period. Our mechanism analyses show that, after comment moderation, contributors may become more conservative in providing new answers by writing more similar, specific, and longer content. In addition, comment moderation exerts a negative impact on the volume of the subsequent regular answers but has no significant effect on harmful answers, which provides supporting evidence for the plausibility of a chilling effect.

Keywords: Platform-initiated comment moderation, Social media platform, Chilling effect, Differences-in-differences.

1. Introduction

Social media has revolutionized the way people communicate, share information, and engage with one another globally. With an estimated 3.96 billion active social media users worldwide as of January 2021, these

platforms have become integral to modern society's fabric.¹ However, alongside social media's undeniable benefits to connect people and facilitate free speech, social media platforms have also encountered significant challenges, including the proliferation of misinformation, hate speech, cyberbullying, and other harmful content. As a result, there has been a growing emphasis on social media moderation, with platforms investing increasingly in content moderation mechanisms and policies. According to Statista, Meta and TikTok remove millions of pieces of content every quarter that violates their community standards.² A growing body of literature has studied the direct impact of content moderation on reducing violations of the moderated content or users. For example, Srinivasan et al. (2019) show that comment removal improves the behavior of the comment author to comply with the community rules. Zhang et al. (2023) demonstrate that the users banned by the platform (for the first time) would generate more content after lifting the ban.

Though prior literature has focused on the direct impact of content moderation, limited attention has been given to its indirect and spillover effects on other content generation, particularly unmoderated content. Social media moderation may exert a spillover effect on unmoderated content through the potential chilling effect on content contributors. The chilling effect is the deterrence of lawful activities that the law did not intend to target (Stoycheff et al. 2019). The chilling effect was first introduced in the law theories and received its first comprehensive exploration in First Amendment jurisprudence (Schauer, 1978). Researchers later investigate the chilling effect online, such as Wikipedia

¹ <https://www.hootsuite.com/newsroom/press-releases/more-than-half-of-the-people-on-earth-now-use-social-media>

² <https://www.statista.com/topics/11495/social-media-content-moderation-and-removal/#topicOverview>

use and Internet search behavior under government surveillance.

Social media moderation may change people's perceived norms and induce fear of negative outcomes even though they engage in unmoderated activities (Matias, 2019; Roloff and Cloven, 1990). As a result, unmoderated activities may be affected. To our knowledge, scarce research has focused on the chilling effect of social media moderation. Social media moderation on one type of content may influence the values other groups receive, resulting in unintended consequences on other groups' behavior. In this study, we attempt to fill the literature gap by studying the effect of comment moderation on the subsequent answer generation on a large Q&A platform. We choose subsequent answer generation because users are more active in commenting on answers than on other type of content like questions. In our research context, 98.31% of the comments are for answers. We aim to answer the following research questions:

1. Does content moderation on social media platforms exert a spillover effect on answer generation? If so, how does it affect the quantity and quality of the subsequent answers?

2. What is the underlying mechanism for the effect of comment moderation on answer generation? How do different types of answers and questions moderate such spillover effect?

We randomly selected 7,969 questions posted before the temporary shutdown of the commenting function on Zhihu.com, China's largest Q&A platform, and tracked all activities relating to them, including answers and user information for both the selected questions and corresponding answers before, during, and after the commenting function shutdown (and restoration). Besides blocking the commenting function, the platform also hid existing comments during the shutdown period. To identify the causal spillover effect, we employ a differences-in-differences (DID) design combined with propensity score matching (PSM).

Our analysis reveals that comment moderation may induce a decrease in the volume of the subsequent answers, with an improved quality during and after the shutdown period. We explain the results from the perspective of a chilling effect on contributors aroused from comment moderation. Our mechanism analyses provide supporting evidence. We show that comment moderation exerts a positive impact on the similarity, concreteness, and length of the subsequent answers, which implies that contributors may have become more conservative in providing new answers by writing more similar, specific, and longer content. In addition, comment moderation exerts a negative impact on the volume of the subsequent regular answers but has no significant effect on harmful answers, which provides

supporting evidence for the plausibility of a chilling effect. Lastly, the spillover effect of comment moderation is more prominent for answers replying to social-oriented questions compared to those replying to professional-oriented questions, implying that answers to social-oriented questions are more susceptible to the influence of a chilling effect.

2. Literature

Our study is related to two strands of literature: social media moderation and the chilling effect. Research on social media moderation investigates the efficacy and efficiency of social media moderation policies. Based on the moderation's target, social media moderation can be classified into user and content moderation (Jiménez Durán, 2022). User moderation directly regulates users, such as providing users with certain information or restricting users' freedom of participation. Matias (2019) shows that making community rules visible to newcomers reduces online harassment and unruly behaviors. Zhang et al. (2023) find that after the user ban is lifted, the previously banned users generate more but lower quality content. Content moderation targets and processes undesirable content. He et al. (2021) demonstrate that algorithm-based content moderation tools can augment human volunteer moderators by stimulating them to moderate more posts. Srinivasan et al. (2019) show that beyond the positive effects of shielding a community from undesirable content, comment removal can also improve the behavior of the comment's author by reducing immediate non-compliance rates. The current research primarily focuses on the direct effect of social media moderation. However, the understudied spillover effect of social media moderation may result in unintended, potentially undesirable outcomes on unmoderated content or users. Mudambi et al. (2023) find that content moderation policies aimed at decreasing the volume and impact of misinformation can result in a spillover of misinformation to other topically related spaces. In contrast to Zhang et al. (2023) and Srinivasan et al. (2019), which focus on user moderation, we examine content moderation, i.e., the spillover effect of social media moderation on unmoderated content.

The chilling effect stems from law theories and became prominent in the United States during the Cold War. The "chilling effects doctrine," a legal doctrine in First Amendment jurisprudence, encouraged courts to treat rules or government actions that "might deter" the free exercise of First Amendment rights "with suspicion" (Penney 2016). Researchers later examine the chilling effect in the online context, such as Wikipedia use and Internet search behavior under

government surveillance. Matthews and Tucker (2017) find that the PRISM revelations most negatively affected online search terms deemed both personally and government-sensitive. Wang et al. (2023) examine the impact of computer misuse acts (CMA) enforcement on users' contributions to cybersecurity relevant topics that are not targeted by the law. They find that CMA enforcement reduces the quantity and the extent of relevance to cybersecurity in discussions in hack forums. Unlike the prior literature that mainly studies the chilling effect of government surveillance, we investigate the chilling effect of social media moderation. To the best of our knowledge, our study is among the first to dive deep into the mechanism of the chilling effect of social media moderation.

3. Research context and data

3.1. Comment shutdown in Zhihu

To empirically examine our research question, we use Zhihu.com, the equivalent of Quora in China, as our research context. Zhihu, launched on January 26, 2011, is a high-quality question-and-answer (Q&A) content platform. Zhihu has some unique advantages to be our research context. First, Zhihu is one of the largest Q&A platforms, with an average of 103 million monthly active users.³ By the end of 2021, Zhihu had a total of 490 million pieces of content in multiple formats, including 420 million Q&As. Second, Zhihu has a commenting function, where users can interact with others by providing comments for an answer or a question. Users are more active in commenting on answers than on questions. 98.31% of the comments are for answers. The commenting function allows us to study the impact of comment-related events on other types of content. Third, the platform launched a comment regulation policy without notifications. Such platform-initiated, external shock provides a base for studying the policy impact on content generation.

On December 20, 2021, the Beijing Internet Information Office requested Zhihu to rectify the content on the platform, as certain information is subject to violate the laws or regulations.⁴ Responding swiftly, on the same day, Zhihu shut down the commenting function platform-wide.⁵ During the shutdown, users were not able to engage in the commenting function (e.g., writing or upvoting) in comment areas, and the existing comments were not available. Zhihu launched the shutdown without prior notification or explanation to users. One week later, Zhihu restored the commenting function, and users could engage in comment activities

again. During the one-week shutdown period, Zhihu has performed further moderation on the existing comments. Content moderation, in general, may include removing the contents that are deemed harmful by the platform or limiting the dissemination of those harmful contents. However, in our context, the specific further moderations the platform performed on the comments has not been explained to the public or users on the platform. Since the temporary commenting function shutdown is equivalent to the temporary content removal, we define content moderation in our research context as the joint action of commenting function shutdown and further undisclosed content moderation.

3.2. Comment shutdown in Zhihu

Our main dataset encompasses two parts, the treatment group and the control group, based on the question posting date. For the treatment group, we randomly selected 7,969 questions posted one week before the temporary shutdown of the commenting function (December 20, 2021) on Zhihu.com. The 7,969 questions were posted on a date between December 13, 2021 and December 19, 2021, i.e., the pre-treatment period. Our data collection includes all activities on the 7,969 questions, their corresponding answers, and user information for both the questions and their corresponding answers. The questions we tracked cover a variety of categories, such as technology, entertainment, and society. We also recorded all activities relating to the 7,969 questions for four weeks during and after the shutdown of the commenting function until January 18, 2022. Because the temporary shutdown affected all users, we constructed the control group by choosing questions posted earlier than the policy implementation period. Similar constructions of control groups are used in prior research (Dewan et al. 2017; Zhao et al. 2023).

Our control group can serve as a counterfactual group for the following reasons. First, user engagements (such as answer and comment writing) in posted questions about similar topics on Zhihu follow a similar pattern. Additionally, we randomly sampled control group questions in the same categories that appeared in the treatment group. Second, the implementation of comment moderation by the platform was exogenous to users, as Zhihu did not provide any forewarnings or hindsight explanations for the moderation. Third, our parallel trend test shows that the treatment and control samples exhibited a common trend prior to the policy (See Appendix A). To test the robustness of our result, we also construct an alternative control group from a

³ <https://app.mokahr.com/apply/zhihu/78336#/>

⁴ <https://mp.weixin.qq.com/s/BwzzovxT7A83g35c86gvDA>

⁵ https://www.sohu.com/a/510393453_120000222

more recent period, i.e., one month before. The results are consistent with our main results (See column (1) of Appendix B).

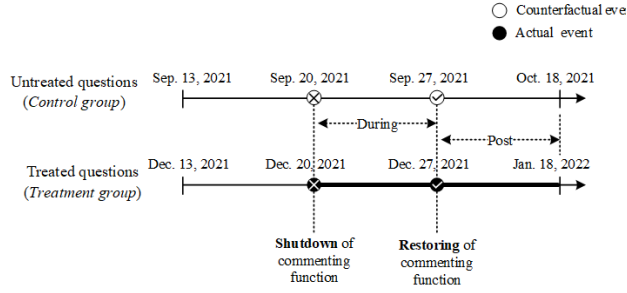


Figure 1. The timeline of the comment moderation event and counterfactual event

Specifically, we randomly selected 7,970 questions posted between September 13, 2021, and September 19, 2021, three months before the shutdown, as our control group. Similar to the treatment group, we record all activities relating to the control group questions for four weeks until October 18, 2021. The specific timeline is shown in Figure 1. In addition, we collected user-level data to ensure the robustness of our analyses. We randomly chose 9,007 users who engaged in answer generation relating to the questions in the treatment and control groups and recorded all their activities (both

relating to and unrelated to the questions in the treatment and control groups) from December 13, 2021, to January 18, 2022.

To investigate the consequences of comment moderation on answer generation, we focus on two dimensions of answer contribution: answer quantity and quality. According to literature (Burtch et al., 2022), the quantity of UGC can reflect users' content-generation motivation and serve as a performance metric. We measure answer quantity by the number of answers posted for a question on a given day, *AnsNum*. The answers that receive more upvotes can indicate their popularity and peer recognition of the content quality (Wang et al., 2022). We use the number of upvotes for an answer to measure the answer quality, *AnsVote*. The logarithmic transformation is performed to address the skewness of these variables (Zhao et al., 2021).

We include a set of control variables to account for the potential confounding factors: (1) The number of followers to a question on a given day, *FollowerNum*; (2) The number of comments to a question on a given day, *ComNum*. (3) The number of invitations to engage in a question on a given day, *InvNum*. Users can invite other users to engage in contributing answers to a question. The summary statistics and variable descriptions are shown in Table 1.

Table 1. Summary statistics of key variables

Variables	Full Sample (N=303,001)				Pre vs. During (N=143,533)				Pre vs. Post (N=216,420)			
	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max
$\text{Ln}(AnsNum_{it})$	0.083	0.307	0	6.486	0.161	0.422	0	6.203	0.082	0.309	0	6.486
$\text{Ln}(AnsVote_{it})$	0.189	0.509	0	8.753	0.213	0.534	0	6.427	0.189	0.505	0	8.753
$ComMod_i \times After_{it}^{(during)}$					0.262	0.44	0	1				
$ComMod_i \times After_{it}^{(post)}$									0.393	0.488	0	1
$\text{Ln}(FollowerNum_{it})$	1.384	0.952	0	8.136	1.29	0.866	0	7.814	1.393	0.966	0	8.136
$\text{Ln}(ComNum_{it})$	0.037	0.21	0	5.333	0.031	0.188	0	4.718	0.038	0.214	0	5.333
$\text{Ln}(InvNum_{it})$	4.512	2.607	0	12.651	4.367	2.549	0	12.625	4.512	2.616	0	12.651

Note: $\text{Ln}()$ indicates that the logarithm is taken for that variable to reduce the skewness and yield elasticity-based interpretation (Zhao et al. 2022). Note that there are 56,952 obs. in the *pre* stage. So, the equivalence among the three sets of samples should be $303,001+56,952 = 143,533+216,420$. The obs. of $\text{Ln}(AnsVote_{it})$ for *Pre vs. During* and *Pre vs. Post* are 18,563, and 14,262 respectively.

4. Empirical analysis and results

4.1. Empirical strategy

We are interested in investigating the impact of comment moderation on the subsequent answer generation. The analyses are performed at the question-day level. We employ a differences-in-differences (DID) estimation strategy, leveraging the comment shutdown on the platform as an exogenous shock. We

use the questions three months before the event as a control group. The DID specification is shown as follows.

$$Y_{it} = \beta_0^{(m)} + \beta_1^{(m)} ComMod_i \times After_{it}^{(m)} + \beta_2^{(m)} ComMod_i + \beta_3^{(m)} After_{it}^{(m)} + Controls_{it} + \alpha_i + \tau_t + \varepsilon_{it} \quad (1)$$

where i denotes question and t denotes day. Y_{it} refers to the outcome variables, including answer quantity $\text{Ln}(AnsNum_{it})$ and answer quality $\text{Ln}(AnsVote_{it})$. The interaction term $ComMod_i \times After_{it}^{(m)}$ is the key independent variable. $ComMod_i$ indicates group

information and takes one if question i is in treated group, and zero otherwise. $After_{it}^{(m)}$ is the treatment period variable which takes one if question i on day t is in the m period (where m includes *during* and *post* period of comment shutdown), and zero before the shutdown. $\beta_1^{(m)}$ is the coefficient of interest, where $\beta_1^{(during)}$ and $\beta_1^{(post)}$ capture the effects of comment moderation during and after the shutdown, respectively. $Controls_{it}$ denotes control variables shown in Table 1. Finally, we use α_i to capture any unobserved question-specific fixed effects and τ_t to control for any unobserved time-specific fixed effects. ε_{it} denotes the error term.

Since there may be systematic differences between the questions in the treatment and control groups, we employ a propensity score matching (PSM) method to account for observed time-variant and time-invariant characteristics. Five covariates before the commenting function shutdown are selected to perform propensity score matching, including: (1) the number of answers to a question; (2) the average number of question page views; (3) the average number of followers to a question; (4) The average number of comments to a question; and (5) the average number of invitations to provide answers for a question. We match the treated and untreated questions with the one-to-one nearest neighbor setting without replacement and with a caliber of 0.01. By conducting PSM, we matched the 5,564 treated questions to 5,881 untreated questions. The balancing test after matching shows no significant differences in the means of all covariates.

4.2. Main results

Table 2 presents the main estimation results. Column (1) of Panel A shows that comment moderation leads to a 2.1% significant decrease in answer quantity during comment shutdown. Column (1) of Panel B presents that, after restoring the commenting function, there is an average 4% reduction in answer volumes in the subsequent three weeks, suggesting that comment moderation may induce detrimental effect on users' engagement in writing new answers. In addition, results in column (2) show that comment moderation may have a positive impact on the quality of the subsequent answers generated. Together, our results suggest that the platform-initiated, comment moderation may exert a significant spillover effect on the subsequent answer generation, which is the unmoderated content. We explore the underlying reasoning in the mechanism analysis in Section 5.

Table 2. The main results

	Answer quantity	Answer quality
	(1)	(2)
	Ln(<i>AnsNum_{it}</i>)	Ln(<i>AnsVote_{it}</i>)
Panel A: Pre vs. During		
<i>ComMod_i</i> × <i>After_{it}</i> ^(during)	-0.021*** (0.005)	0.034* (0.02)
Obs.	143,533	18,563
Panel B: Pre vs. Post (Three weeks)		
<i>ComMod_i</i> × <i>After_{it}</i> ^(post)	-0.04*** (0.008)	0.091*** (0.001)
Obs.	216,420	14,262
Controls	Y	Y
Post FE	Y	Y
Day FE	Y	Y

Notes: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors are clustered at post level. As some questions did not receive any answers, the number of observations in column Ln(*AnsNum_{it}*) is smaller than that in column Ln(*AnsVote_{it}*).

4.3. Robustness check

4.3.1. Relative time models. A main assumption of DID design is the parallel trend assumption, which describes that the variations in the outcome of the control group should exhibit a similar pattern as that of the treatment group prior to the external shock. In other words, the change of the control group may serve as a counterfactual change for that of the treatment group if there were no treatments. To empirically test this assumption, we include time dummies indicating seven days before and 24 days after the commenting function shutdown (Angrist and Pischke, 2008). The regression specification is as follows.

$$Y_{it} = \beta_0 + \sum_{-7}^{28} \delta_{\omega} D_{i\omega} + Controls_{it} + \alpha_i + \tau_t + \varepsilon_{it}, \quad (2)$$

where $D_{i\omega}$ are a series of pre-moderation time dummies that denote the chronological distance between the observation time and the comment moderation start time. δ_{ω} are the coefficients that capture the comment moderation-induced changes over time.

We present the result for the outcome variable Ln(*AswNum_{it}*) in Appendix A. We find that the pre-moderation coefficients are near zero and not statistically significant, which suggests that the answer quantity of the treated and untreated questions are comparable prior to the introduction of the comment moderation policy. This supports a parallel trend regarding answer generation between the two groups.

4.3.2. Alternative control group: One month before shutdown. An alternative control group is a set of questions closer to the moderation policy. We randomly select questions created from November 8, 2021, to November 8, 2021, one month before the shutdown. The observation window for these questions lasts until December 14, 2021, again four weeks after the shutdown. We use the alternative sample and redo the regression in equation (1). The estimation results are shown in Column (1) of the Appendix B. The results remain consistent with our main results.

4.3.3. Change in user-level activity. We further test the robustness of our results by examining the changes in answer contributions at the user-level. We perform an estimation on the answer generations of the 9,007 users, who engaged in question or answer generation relating to the treatment and the original control group, before and after the shutdown. The results presented in Column (2) of Appendix B show that, after moderation, the volume of user answers exhibits a decrease pattern. This decline persists even after the shutdown, which is consistent with our main results.

5. Mechanism analysis

First, we discuss the chilling effects of users as a potential theoretical explanation for the underlying mechanism. Then we empirically test this explanation by examining several possible channels related to the changes in the answers on the platform.

Chilling effects pertain to how the fear of negative consequences may cause individuals to avoid participating in legal activities at which the regulation was not intended to target (Rolloff and Cloven, 1990; Wang et al., 2023). In our context, the platform-initiated comment moderation was not transparent to users. Specifically, the platform did not disclose the policy details in terms of what contents in comments have been regulated and whether certain forms of penalty have been applied for the associated users. The opacity of such moderation in comments may raise user concerns and fears. Users might be hesitant to participate in writing new answers even if answers were not moderated. Those who participate might be more cautious and care more about the quality of their answers. Hence, a chilling effect may be aroused from comment moderation and cause a decline in answer volume and, meanwhile, a positive change in the quality of the subsequent answers.

5.1. Answer similarity, concreteness, and length

To test this theoretical explanation, we dive into the changes in the content of answers to seek empirical evidence. We develop three measurements: answer similarity, concreteness, and length. First, answer similarity is defined as the closeness of the newly generated answer with all existing answers to the same questions (Deng et al., 2022). In the formula below, $AnsSim_{it}$ is the average of the similarities of all new answers to question i on day t with all existing answers to the same question in previous days. We denote the vector of each answer α to question i on day t as V_{it}^{α} . All answers to question i from day 1 over day $t-1$ are represented as one vector $V_{i,[1,t-1]}$.

$$AnsSim_{it} = \frac{\sum_{\alpha=1}^N \text{Cos}(V_{it}^{\alpha}, V_{i,[1,t-1]})}{N}$$

Second, linguistic concreteness assesses how effectively words convey descriptive, specific, and vivid details about an object or situation (Hansen and Wänke 2010). We measure the answer concreteness, $\text{Ln}(AnsCon_{it})$, using a well-developed dictionary to calculate the concreteness score embedded in the answer text (Xie and Bi 2022). A higher value indicates more concrete content. Third, we use the word count (in Chinese characters) to measure the answer length, $\text{Ln}(AnsLen_{it})$ (Burtch et al., 2022).

Using the three indicators as dependent variables, we perform the regression with equation (1) and present the estimation results in Table 3. We find that the subsequent answers tend to be more concrete and similar to the existing ones. The increased similarity implies that, after comment moderation, users may behave more conservatively by providing perspectives similar to the existing answers. In the meantime, users may provide concrete answers to reduce ambiguity and risk of being misunderstood. These results suggest that comment moderation exerts a chilling effect on the answer contributors. As presented in Column (3) of Table 3, the positive effect of comment moderation on the answer length provides supporting evidence that new answer contributors make more effort in writing the answers to reduce the likelihood of being misinterpreted and regulated. Therefore, the chilling effect of comment moderation may cause an improvement in the quality of the subsequent answers.

Table 3. Impacts of comment moderation on answer similarity, concreteness, and length

DV	(1) <i>AnsSim_{it}</i>	(2) <i>Ln(AnsCon_{it})</i>	(3) <i>Ln(AnsLen_{it})</i>
Panel A: Pre vs. During			
<i>ComMod_i</i> × <i>After_{it}^(during)</i>	0.013*** (0.003)	0.109** (0.049)	0.092** (0.047)
Obs.	18,563	18,563	18,563
Panel B: Pre vs. Post (Three weeks)			
<i>ComMod_i</i> × <i>After_{it}^(post)</i>	0.008** (0.004)	0.124** (0.063)	0.119** (0.006)
Obs.	14,262	14,262	14,262
Controls	Y	Y	Y
Post FE	Y	Y	Y
Day FE	Y	Y	Y

Notes: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors are clustered at post level.

5.2. Regular answers and harmful answers

We further explore how the impact of comment moderation may differ across regular answers and

harmful answers. To detect the answer harmfulness, we use Baidu Text Censor API to detect the harmful content (Zhang et al., 2023). This API can identify whether the focal text is harmful and, if so, which type of harmful content. After using this detector, 7,636 harmful answers (out of 109,147) are detected, accounting for 6.996% in our sample. The distribution of the specific answer types is shown in Appendix C. A large proportion of harmful answers are those related to marketing, abusive, and political content.

We measure the effect of comment moderation on the regular answers and the identified harmful answers. The results are in Table 4, which shows that comment moderation induces a decrease in the volume of regular answers but not in the volume of harmful answers, except for a slight reduction in the volume of advertisement answers. These findings provide indirect evidence of the chilling effect, i.e., the content regulation may demotivate respondents from providing new regular answers. Meanwhile, we show that the comment regulation does not exert a significant effect on the harmful answers. Therefore, our results send a caveat to UGC platforms that the good-intent content moderation policy may result in unintended consequences by suppressing regular content but not the production of harmful content.

Table 4. Heterogeneity across answer type

DV	Regular answers	Abusive answers	Political answers	Advertisement answers	Porn answers	Answers with banned_words	Violent answers
Panel A: Pre vs. During							
<i>ComMod_i</i> × <i>After_{it}^(during)</i>	-0.02*** (0.005)	-0.001 (0.001)	-0.001 (0.001)	-0.0003 (0.001)	-0.0001 (0.001)	-0.0001 (0.002)	-0.0001 (0.0001)
Panel B: Pre vs. Post (Three weeks)							
<i>ComMod_i</i> × <i>After_{it}^(post)</i>	-0.034*** (0.008)	-0.0004 (0.001)	-0.001 (0.001)	-0.002* (0.001)	0.0001 (0.001)	5.89e-06 (0.0003)	-0.00001 (0.0002)
Controls	Y	Y	Y	Y	Y	Y	Y
Post FE	Y	Y	Y	Y	Y	Y	Y
Day FE	Y	Y	Y	Y	Y	Y	Y

Notes: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors are clustered at post level.

5.3. Social-oriented and professional-oriented questions

To further examine the plausibility of the chilling effects explanation, we study how the impact of comment moderation may differ for answers to different types of questions. The topicality of a question is dependent on what categories the question

belongs to (Chae et al., 2017). The Zhihu platform often tags one or more categories for a question. We use the first tag as the category of a question. There are 31 categories in our dataset in total. We divide all the categories into two dimensions, i.e., professional and social. The professional-oriented questions pertain to objective or knowledge domains, such as engineering, laws, economics, and natural sciences. Answers to these questions are usually more objective and less controversial. Social-oriented questions may contain

personal viewpoints and social interactions among users and are typically related to news, emotions, and society. Answers to these questions are generally more subjective and contentious and might contain prejudicial or even abusive language. Compared to answers to professional-oriented questions, answers to social-oriented questions are more likely to be moderated by the platform and, therefore, are more susceptible to the influence of chilling effects.

Table 5. The heterogeneity in question category

	Profession-oriented questions	Social-oriented questions
	(1)	(2)
	Ln(AnsNum _{it})	Ln(AnsNum _{it})
Panel A: Pre vs. During		
<i>ComMod_i</i>	-0.012	-0.027***
× <i>After_{it}^(during)</i>	(0.008)	(0.006)
Obs.	60,928	82,590
Panel B: Pre vs. Post (Three weeks)		
<i>ComMod_i</i>	-0.012	-0.06***
× <i>After_{it}^(post)</i>	(0.013)	(0.011)
Obs.	91,901	124,508
Controls	Y	Y
Post FE	Y	Y
Day FE	Y	Y

Notes: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors are clustered at post level.

We run equation (1) using the subsamples corresponding to the two question categories. As shown in Table 5, the decrease in answer volume mainly occurs for social-oriented questions rather than professional-oriented questions. This result suggests that comment moderation may exert a chilling effect and demotivate the answer generation for social-oriented questions. Such negative effect can be detrimental to the development and sustainability of the social community.

5.4. Casual mediation analysis through page view

Our studied Q&A content platform is a multi-sided market that serves multiple distinct sides, whose ultimate benefit stems from interacting through a common platform (Rochet and Tirole 2003). The users seeking content can be regarded as the demand side, and those who generate content can be viewed as the supply side. More demand stimulates more supply and vice versa in a Q&A content platform. Shutting down

the commenting function blocks the supply of comments, a type of content, which may result in a decrease in demand, which can be measured by the number of page views of a question. With less demand, users are less incentivized to provide more content, such as answers. We use the casual mediation analysis to investigate whether page view may mediate the relationship between comment moderation and answer generation. The estimated equations are as follows.

$$(1) E[M|a, c] = \beta_0 + \beta_1 a + \beta_2 c$$

$$(2) E[Y|a, m, c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_4 c$$

$$(3) E[Y|a, m, c] = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4 c$$

where a is independent variable (i.e., comment moderation). m is mediator (i.e., Δ page view). Y is outcome variable (i.e., Ln(AnsNum_{it})). c is the control variables, as with equation (1). The results are the estimates of the controlled direct effects, the natural indirect effect, and the natural direct effect.

The estimated results are shown in Table 6. The result shows that page view significantly mediates the negative effects of comment moderation on answer quantity. That is, the comment moderation induces a reduction in users' needs to view platform content. Less demand for content, in turn, lowers users' enthusiasm for answer contributions.

6. Discussion and conclusion

Despite a growing body of literature examining the direct impact of social media moderation on reducing harmful behavior, limited attention has been given to its spillover effect on other content, particularly on unmoderated content. Exploiting a temporary shutdown of the commenting function on a large Q&A platform, our study is among the first to examine the spillover effect of comment moderation on the subsequent answer generation. We quantify the spillover effect by comparing answer quantity and quality before and during/after the shutdown.

We find that the comment ban has detrimental impacts on user engagement in generating answers but positively affects the quality of the answers generated. Delving deeper into mechanism, we find that the result can be explained from the perspective of a chilling effect, where the fear of negative consequences prevents users from generating answers that the comment moderation was not intended to target. We examine the chilling effect through three channels: the content of answers (measured by the similarity between answers, answer concreteness, and the length of answers), the types of answers (regular vs. harmful answers), and the types of questions answers reply to (social-oriented vs. professional-oriented questions).

Our results show that the subsequent answers tend to be similar to the existing ones but longer and more concrete, implying that users may behave more conservatively and make more effort in writing answers so as to reduce ambiguity and the risk of being misunderstood. This may explain why comment moderation induces an improvement in answer quality. In terms of the types of answers affected by the spillover effect, we find that regular rather than harmful answers are negatively affected, which provides further evidence of the chilling effect of comment moderation. As for the heterogeneous effects that comment moderation has on answers to different types of questions, we show that compared to professional-oriented questions, answers to social-oriented questions, which tend to be more subjective

and contentious, are more susceptible to the influence of chilling effects.

Our results demonstrate that comment moderation is a double-edged sword in that answer quality improves, but answer quantity reduces. The chilling effect of comment moderation on answers to social-oriented questions may harm the development and sustainability of online social communities. Social media platforms need to assess the direct and indirect benefits and damages before initiating social media moderation, as moderation may have a chilling effect on desirable content generation. Future research can examine comment moderation on question generation or other content moderation on the unmoderated content.

Table 6. The casual mediation analysis of page view

	Pre vs. during	Pre vs. post
Mediation effects: content mod. $\rightarrow \Delta$ page view $\rightarrow \text{Ln}(AnsNum_{it})$	-0.006***	-0.003***
Direct effects: content mod. $\rightarrow \text{Ln}(AnsNum_{it})$	-0.017***	-0.034***
Total effects: content mod. on $\text{Ln}(AnsNum_{it})$	-0.023***	-0.037***
Proportion mediated	0.261	0.081

References

- Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.
- Burch, G., He, Q., Hong, Y., & Lee, D. (2022). How do peer awards motivate creative content? Experimental evidence from Reddit. *Management Science*, 68(5), 3488-3506.
- Chae, I., Stephen, A. T., Bart, Y., & Yao, D. (2017). Spillover effects in seeded word-of-mouth marketing campaigns. *Marketing Science*, 36(1), 89-104.
- 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.
- Dewan, S., Ho, Y. J., & Ramaprasad, J. (2017). Popularity or proximity: Characterizing the nature of social influence in an online music community. *Information Systems Research*, 28(1), 117-136.
- Hansen J, Wänke M (2010) Truth from language and truth from fit: The impact of linguistic concreteness and level of construal on subjective truth. *Personality and Soc. Psych. Bulletin* 36(11):1576-1588.
- He, Q., Hong, Y., & Raghu, T. (2021). The Effects of Machine-powered Platform Governance: An Empirical Study of Content Moderation. Available at SSRN.
- Jiménez Durán, R. (2022). The economics of content moderation: Theory and experimental evidence from hate speech on Twitter. Available at SSRN.
- Li, X., Grahl, J., & Hinz, O. (2022). How do recommender systems lead to consumer purchases? A causal

- mediation analysis of a field experiment. *Information Systems Research*, 33(2), 620-637.
- Marthews A, Tucker C (2017) Government surveillance and Internet search behavior. Preprint, submitted February 17, <http://dx.doi.org/10.2139/ssrn.2412564>.
- Matias, J. N. (2019). Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences*, 116(20), 9785-9789.
- Mudambi, M., Clark, J., Rhue, L., & Viswanathan, S. (2023). Fighting Misinformation on Social Media: An Empirical Investigation of the Impact of Prominence Reduction Policies. Available at SSRN 4653460.
- Peng, J. (2023). Identification of causal mechanisms from randomized experiments: A framework for endogenous mediation analysis. *Information Systems Research*, 34(1), 67-84.
- Penney JW (2016) Chilling effects: Online surveillance and Wikipedia use. *Berkeley Tech. Law J.* 31:117-182.
- Rochet, J. C., & Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4), 990-1029.
- Roloff ME, Cloven DH (1990) The Chilling Effect in Interpersonal Relationships: The Reluctance to Speak One's Mind. The International Communication Association, Dublin, Ireland (Lawrence Erlbaum Associates, Mahwah, NJ).
- Schauer, F. (1978). Fear, risk and the first amendment: Unraveling the chilling effect. *BUL rev.*, 58, 685.
- Srinivasan, K. B., Danescu-Niculescu-Mizil, C., Lee, L., & Tan, C. (2019). Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on human-computer interaction*, 3(CSCW), 1-21.

Stoycheff E, Liu J, Xu K, Wibowo K (2019) Privacy and the panopticon: Online mass surveillance’s deterrence and chilling effects. *New Media Soc.* 21(3):602–619.

Wang, J., Li, G., & Hui, K. L. (2022). Monetary incentives and knowledge spillover: Evidence from a natural experiment. *Management Science*, 68(5), 3549-3572.

Wang, Q. H., Geng, R., & Kim, S. H. (2023). Chilling Effect of the Enforcement of Computer Misuse Act: Evidence from Publicly Accessible Hack Forums. *Information Systems Research*.

Xie Z, Bi R (2022) Construction and inference technique of large-scale Chinese concreteness lexicon. *Acta Scientiarum Naturalium Universitatis Pekinensis* 58(1):1–6.

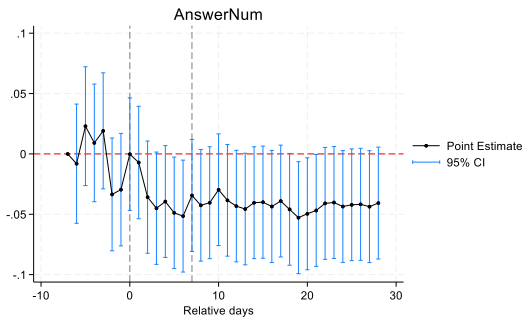
Zhang, X., Wei, Z., Du, Q., & Zhang, Z. (2023). Social media moderation and content generation: evidence from user bans. Available at SSRN 4089011.

Zhao, W., Liu, Q. B., Guo, X., Wu, T., & Kumar, S. (2022). Quid pro quo in online medical consultation? Investigating the effects of small monetary gifts from patients. *Production and Operations Management*, 31(4), 1698-1718.

Zhao, K., Lu, Y., Hu, Y., & Hong, Y. (2023). Direct and indirect spillovers from content providers’ switching: Evidence from online livestreaming. *Information Systems Research*, 34(3), 847-866.

Appendix

Appendix A. The relative time model



Appendix B. Robustness check

	(1) Alternative control group	(2) User-level: # of answer
Panel A: Pre vs. During		
$ComMod_i$ $\times After_{it}^{(during)}$	-0.52*** (0.025)	-0.045*** (0.006)
Obs.	111,127	125,368

Panel B: Pre vs. Post (Three weeks)		
$ComMod_i$ $\times After_{it}^{(post)}$	-0.26*** (0.005)	-0.046*** (0.006)
Obs.	332,532	260,348
Controls	Y	Y
Post FE	Y	Y
Day FE	Y	Y

Notes: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Robust standard errors are clustered at post level and user level, respectively.

Appendix C. The distribution of harmful answers

