The Effects of Diversity in Algorithmic Recommendations on Digital Content Consumption: A Field Experiment

Guangying Chen,[*] Tat Y. Chan,[*] Dennis J. Zhang,[*] Senmao Liu,[†] Yuxiang Wu[†]

February 20, 2023

Abstract

Social media platforms such as TikTok and Facebook are criticized for trapping consumers in "filter bubbles" (Pariser 2011) through personalized recommendations based on users' detailed individual information. Practitioners and regulators have been calling for platforms to tackle the problem by incorporating more diversified content in their recommender systems. We aim to study the causal effects of more diversified personalized recommendations on users' behaviors in practice. By collaborating with NetEase Cloud Music, the world's third-largest music-streaming service company, we developed a new recommender system with more content diversity based on their existing state-of-the-art recommender system. We then conducted a large-scale field experiment where hundreds of millions of users were randomly assigned to receive video recommendations either from the platform's current recommender algorithm or our modified algorithm. Although the new algorithm increased the diversity of recommended content to users, overall there is no clear evidence that it increased the diversity of consumed content, but it decreased users' consumption level. However, for active users, we find that a 1% increase in recommendation diversity boosted their consumption diversity by 0.55% without reducing their consumption or engagement level. We show that the accuracy of predicting users' preferences is key for the new algorithm to increase the consumption diversity and, when users also highly value the platform, their consumption and engagement levels will not be hurt. The company eventually adopted our algorithm modification and now uses it to serve millions of customers daily.

Keywords: Recommender Systems, Filter Bubble, Social Media Platforms, Recommendation Diversification, Field Experiment

[*] Guangying Chen (email: guangyingchen@wustl.edu), Tat Y. Chan (email: chan@wustl.edu), and Dennis J. Zhang (email: denniszhang@wustl.edu): Olin Business School, Washington University in St. Louis, St. Louis, MO, USA.
[†] Industry collaborators: NetEase Cloud Music, Inc., Hangzhou, Zhejiang, China.

1. Introduction

Social media is ubiquitous. In January 2023, 59.4% of the world's population used social media, on average 2.5 hours every day.[1] Almost all large social media platforms, including TikTok and Facebook, personalize content recommendations based on users' individual historical activity to entice users to spend more time on the platform—and generate more revenue through digital advertising. Such personalized recommender systems often trap consumers in their own "filter bubbles" (Pariser 2011): platforms recommend only content with opinions and information that conform to users' existing beliefs. Many observers have noted that this behavior may severely enhance an individual's bias.

Practitioners and regulators have been calling for social media platforms to tackle the problem by incorporating more diversified content in their recommender systems. For example, the U.S. Senate introduced the Filter Bubble Transparency Act (S.2024) in June 2021, requiring online platforms to allow consumers to conveniently opt out of personalized recommendations.[2] As mentioned by the bill's sponsor, John Thune, the act aims to give consumers an option "to see information that has not been selected specifically for them" and help them expand their consumption diversity.[3]

Does disabling personalized recommendations help solve the filter-bubble problem on social media platforms? Evidence shows that it might not. First, a global survey in the Reuters Institute's 2016 Digital News Report[4] revealed that most people prefer to get news from personalized recommendations (36%) compared to editorial/journalistic recommendations (30%) or social recommendations (22%). Second, people may ignore the more diversified content recommended by platforms and stay in their own filter bubble. For example, a Facebook experiment in 2018 showed that disabling the personalized News Feed ranking algorithm led users to hide 50% more recommended posts, mostly due to their lack of interest in the topics. More concerning, users sharply increased their usage of Facebook Groups, which often contain

[1] https://datareportal.com/social-media-users
[2] https://www.congress.gov/bill/117th-congress/senate-bill/2024; https://www.makeuseof.com/filter-bubble-transparency-act-explained/
[3] https://www.thune.senate.gov/public/index.cfm/press-releases?ID=F0D3EA8C-D573-4A7F-89BB-027ABE2781F2
[4] https://www.digitalnewsreport.org/essays/2016/people-want-personalised-recommendations/

more extreme content.[5] From platforms' perspective, diversified recommendations may reduce users' interest, and consequently their content consumption and engagement levels. For example, the Facebook experiment[5] found diversified recommendations lowered the number of comments between friends by 20%. Therefore, it is important for policymakers and social media platforms to understand how diversified recommendations may affect consumers' behavior and how to effectively encourage consumers to consume more diversified content.

Given the importance of diversified recommendations for consumers, platforms, and society, our research tries to address the following questions: (1) How does modifying the current recommender algorithms on social media platforms to recommend more diversified content change users' consumption behavior? (2) Do consumers differ in how they react to more diversified recommendations? (3) How can social media platforms improve their recommendation diversity without hurting important performance metrics such as user consumption and engagement?

To answer these questions, we partnered with NetEase Cloud Music (NCM), the world's third-largest music-streaming service company, to conduct a large-scale, 14-week randomized field experiment in the Cloud Village page of NCM's mobile app. NCM has built a state-of-the-art recommender system to personalize the recommendation of music-related videos for individual users. Unlike past studies in which an advanced personalized algorithm is compared with a more diverse algorithm but rarely used by platforms (e.g., Claussen et al. 2019, Holtz et al. 2020), we modified the existing personalized algorithm to increase video-topic diversity. Doing so allows us to minimize the negative impact on users' consumption and engagement and to make our insights more generalizable and easier to implement for other platforms.

In the experiment, users of the treatment group saw content recommended by our new, more diverse algorithm, and users of the control group saw content recommended by the original system. We found that the new algorithm exposed the treatment group to 2.33% more topic diversity (measured by the Herfindahl–Hirschman Index, HHI) than the control group. Our results show that, overall, the effects of the new

[5] https://bigtechnology.substack.com/p/facebook-removed-the-news-feed-algorithm?s=09

algorithm on users' consumption diversity are mixed; however, the algorithm lowered their weekly clicks by 3.08%, confirming the concerns of social media platforms that diversified recommendations can hurt the business.

We then investigate how users reacted differently to the new algorithm. We find no significant effects on new users' consumption and engagement behavior or on their consumption diversity. For inactive users (those who have been inactive for more than four weeks), the new algorithm only decreased their consumption level without affecting their engagement level or consumption diversity. However, for active users, the new algorithm effectively boosted their consumption diversity—a 1% increase in recommendation diversity led to a 0.55% increase in the consumption diversity—without hurting their consumption and engagement levels. These results suggest that, while the trade-off between consumption and diversity exists on average, higher recommendation diversity can effectively mitigate the filter-bubble problem for active users, who contribute the most viewing time and revenue to the platform, without sacrificing important performance metrics.

We further explore what drives the positive reactions of active users. Our results reveal that, for only the users who spend significant time on the platform *and also* click on a large number of videos, consumption diversity would increase without hurting their consumption and engagement levels. Interestingly, for active users who spend a lot of time on the platform but view only a few (long) videos, the increased recommendation diversity did not increase their consumption diversity but did lower their video clicks, view time, and number of likes and comments left. This result implies that two conditions are crucial for improving the effectiveness of recommendation diversity: (1) when users highly value the platform, and (2) when the company well understands individual consumption preferences through users' click data.

Our research has direct practical impact: based on our findings, NCM updated its algorithm by increasing recommendation diversity to active users. The updated algorithm is used to serve millions of customers daily and has significantly increased users' consumption diversity and consumption level.

Our research builds on marketing and computer science literature that studies the social and economic impacts of algorithmic recommendations, including political affiliations (e.g., Bakshy et al. 2015, Huszár et al. 2022, Ribeiro et al. 2020), product purchase (e.g., Fleder and Hosanagar 2009, Ghose et al. 2014, Lee and Hosanagar 2019, Song 2021), and media consumption (e.g., Berman and Katona 2020, Holtz et al. 2020, Hosanagar et al. 2014, Moehring 2022, Zhou et al. 2010). Our paper contributes to the emerging literature on how diversification of personalized algorithmic recommendations affects the level and diversity of users' digital consumption.

Previous studies mainly compare a personalized recommender algorithm with a less-efficient recommender algorithm, such as popularity-based (Holtz et al. 2020), time-based (Dujeancourt et al. 2022), and human recommendations (Claussen et al. 2019), which are not commonly used by social media platforms nowadays. It has been documented that personalized algorithmic recommendations lead to a higher consumption level. For example, using observational data from an online music service, Hosanagar et al. (2014) found that consumers purchased more songs after receiving personalized recommendations. Claussen et al. (2019) conducted a field experiment on a major German news website and found that personalized algorithmic recommendations generated more user clicks than blanket recommendations from human editors. Holtz et al. (2020) also showed through an online field experiment on Spotify that consumers significantly increased podcast streams after the platform replaced popularity-based recommendations with personalized recommendations. At the same time, personalized recommendations can polarize user consumption and trap users in their filter bubbles. For example, Claussen et al. (2019) and Holtz et al. (2020) found that personalization decreased users' consumption diversity measured by the HHI and Shannon entropy of consumed topics. Claussen et al. (2019) further demonstrated that the declining effect could spread to users' news consumption in other non-personalized sections. Anderson et al. (2020) also ran a field experiment on Spotify and found personalized recommendations performed better for users with lower consumption diversity. Contrary to the prior literature, we demonstrate that making a state-of-the-art recommender system more diversified could affect users' consumption level and consumption diversity. We believe our findings are important because platforms that want to improve their

recommendation diversity are more likely to adjust their existing algorithms instead of using a nonpersonalized algorithm. Our field experiment demonstrates that the modification from the existing algorithm is easy and fast.

Our research also contributes to the literature on recommendation diversification in computer science. Since Bradley and Smyth (2001) first introduced diversification into recommender systems, a large body of research on recommendation diversity has emerged in computer science (Kaminskas and Bridge 2016, Kunaver and Požrl 2017). Most studies focus on measuring the diversity of recommendations (e.g., Clarke et al. 2008, Fleder and Hosanagar 2009, Vargas et al. 2014, Vargas and Castells 2011), evaluating the effects of diversification on recommendation accuracy and consumer satisfaction (e.g., Adomavicius and Kwon 2011a, Ekstrand et al. 2014, Hurley and Zhang 2011, Javari and Jalili 2015), and improving diversification algorithms (e.g., Adomavicius and Kwon 2011b, Vaishnavi et al. 2013, Ziegler et al. 2005). We extend this literature in two ways. First, unlike the previous studies that are based on correlational analysis or surveys, we study the causal effects through a real-world field experiment. Second, we explore how consumers may respond differently to the recommendation diversification.

Finally, our paper provides important practical value for practitioners. First, our results guarantee external validity through a large-scale, randomized field experiment involving hundreds of millions of platform users. Second, since our experiments were implemented in a personalized recommender system currently used in practice, other digital platforms could easily modify the design and incorporate the changes into their own recommender system. Third, we provide suggestions for platforms on how to lift users' consumption diversity without hurting important performance metrics for the business. Specifically, platforms should increase the recommendation diversity for heavy users whose preferences are also well understood by recommender algorithms. By contrast, for new and inactive users, the platform should be cautious in diversifying recommendations. With little knowledge about these users' preferences, adopting a more conservative strategy could help platforms collect more information about their consumption preferences, which can be used for the diversification of recommendations in the future.

2. Field Setting and Experimental Design

NCM had over 800 million users through 2019 and 181.9 million monthly active users[6] in the first half of 2022. Almost all users access NCM services through its mobile app. This app has a main tab called the Cloud Village, where users can watch music-related videos created by the platform's users and recommended by the platform's algorithm (Figure 1, left panel). NCM categorizes each video into one of about 80 topics: music sharing, film mashup, instrumental performance, dancing, and so on. As on other social media platforms, users who watch a video can like, comment on, and share the video (Figure 1, right panel). We implemented our experiment in this Cloud Village.

Figure 1. The Cloud Village Tab (left) and a Music-Related Video (right)

2.1. The Current Personalized Recommender System

The personalized, deep-learning recommender system used by NCM is standard in the industry. Every time a user visits the tab, the system receives a query and will recommend a fixed number of videos from the millions contained in its video pool. The selection algorithm consists of three stages: (1) Retrieval, (2) Ranking, and (3) Re-ranking (see Figure 2 for a simplified example).

[6] https://ir.netease.com/static-files/a1d0942b-2362-49b1-b88e-ab14f7c8359e and https://www1.hkexnews.hk/listedco/listconews/sehk/2022/0922/2022092200787.pdf

(1) **Retrieval:** For a user, the platform uses multiple strategies to retrieve X (around 500 to 10,000) videos from the video pool. The most common strategy is the user-to-item (u2i) model, which accounts for about 40% of video retrievals. Each video candidate (or the user) is represented by an embedding that is pretrained based on the video's features (or the user's historical activities on the platform). The model calculates the distance between the user and each video by passing a sigmoid function of the dot product of user and video embeddings and selects videos with close distances. Such videos are usually from topics the user has clicked on or from a creator the user has liked. Other strategies include cold-start (~10%), item-to-item model (~8%), video popularity (~5%), and hot events (~5%).

(2) **Ranking:** An algorithm then ranks these X videos by their predicted match value with the user and passes a list of top Y (around 100 to 300) videos to the next stage. To predict the match value, the algorithm first trains several deep neural networks to predict the user's behaviors, such as the probability to click or to share and how long the user will watch the video, using the user's and the video's features. The algorithm then calculates the match value as a function of these predictions. Note that the algorithm may rank a popular video very high even though the user has not seen it before.

(3) **Re-ranking:** Another algorithm then re-ranks the Y videos to decide which Z (around 10 to 20) videos to recommend and the listing order. The algorithm first uses deep neural networks to re-predict the match value, adding more complex user-video features, and to re-rank the videos. This step aims to improve the prediction accuracy while saving computational resources by focusing only on a small set of high-value videos. Next, other strategies, including adding advertisements on certain blocks (like every seventh video), list optimization, and the focus of our paper, *content diversification,* will be considered. For platforms such as NCM, content diversification helps users to explore new interests and avoid boredom, which can increase user consumption in the long run. The algorithm adjusts the diversity level of videos through a parameter called window size (denoted as S). The larger the window size, the more likely the user will be recommended more diversified

videos.[7] Specifically, the algorithm first picks the video with the highest match value. For the second recommended video, the algorithm chooses (from the remaining top *S* videos) the most different video compared with the first one based on the topic, creator, and other characteristics. For the third recommended video, the algorithm will choose (from the remaining top *S* videos) the one that is most different from the first and second videos. This process will continue until *Z* videos are selected, which will be recommended to the user in order.

If, after viewing the *Z* recommendations, the user continues to swipe up, the algorithm will remove all previously recommended videos from the pool and repeat the three stages outlined above.

[7] Online Appendix A illustrates why setting a larger window size can effectively increase the content diversity of recommendations. We randomly select a Cloud Village user and simulate the algorithm's recommendations to her query. We ask the algorithm to retrieve 400 videos, select 50 top-ranked videos, and re-rank the 50 videos using window size 5 or 30. We show that the recommendations under window size 5 are less diversified than under 30.

Figure 2 The Current Personalized Recommender System of the Cloud Village: A Simplified Example

2.2. The Experimental Setting

We conducted a large-scale field experiment for 14 weeks from December 22, 2021, to March 29, 2022, randomly sampling 10 million users who visited the Cloud Village at least once during the experiment. Upon the arrival of users, we randomly assigned around 3% to the treatment group and the remaining 97% to the control group, based on a hash function of their user IDs.

Users in the control group were recommended by the original recommender system, in which the window size is 5 for users who have clicked on videos in the past 30 days, and 15 otherwise. The rationale of the platform is that the match value is better predicted for active users with more behavioral data, hence a smaller window size will boost consumption and engagement. In contrast, the prediction for new and inactive users is much less accurate, thus a larger window size could increase the probability that a user will see a video of interest. For users in the treatment group, we modified the algorithm by increasing the window size in the re-ranking stage to 30 for both types of users. This directly increased the likelihood of recommending different content. Recommendation diversity can be increased in other ways as well, such as including additional diverse retrieval sources or changing the ranking procedure, but the implementation would be more complicated and would depend on platforms' existing recommendation strategies.

3. Data

We divide our data into three periods: (1) the four-week pre-experiment period, November 19 to December 16, 2021; (2) the implementation period,[8] December 17 to December 21, 2021; and (3) the 14-week experiment period,[9] December 22, 2021, to March 29, 2022. We use the pre-experiment data to conduct randomization checks and classify users, and we use the experiment data to conduct manipulation checks and estimate the causal effects of recommendation diversity.

[8] NCM increased the window size of treatment-group users to 15 on December 17, 2021, then further increased it to 30 on December 21, 2021.
[9] In Online Appendix B, we fold the implementation period into the experiment period and report the qualitatively consistent results.

We collected users' demographic information, including their app registration date, gender, and age. We tracked their weekly activities in the Cloud Village, such as whether they visited the Village, whether they clicked on at least one video and, if so, how long they watched them, and whether they liked, commented on, or shared them. We also tracked which videos have been recommended to users. Consistent with the literature (Claussen et al. 2019, Holtz et al. 2020), we collected the topic of each video and used three indices to measure the content diversity of videos recommended to or watched by users: the number of unique video topics (num_topic_{it}), the HHI of video topics (HHI_topic_{it}), and the Shannon entropy of video topics ($entropy\_topic_{it}$). The latter two are defined as follows:

$$HHI\_topic_{it} = \sum_{j=1}^{num\_topic_{it}} s_j^2 \quad (1)$$

$$Entropy\_topics_{it} = -\sum_{j=1}^{num\_topic_{it}} s_j \cdot ln\left(s_j\right) \quad (2)$$

where num_topic_{it} indicates the number of unique topics user *i* was recommended or consumed in week *t,* and s_j ($\in (0,1]$) indicates the share of videos of topic j. *The content diversity increases as the number of topics increases, the HHI decreases, and Shannon entropy increases.* Based on Van Dam (2019), the concept of diversity includes not only the "variety" dimension (i.e., the total number of categories in the taxonomy) but also the "balance" dimension (i.e., the distribution of elements across categories). Thus, the HHI and Shannon entropy are more comprehensive indicators of content diversity because they also consider how evenly distributed the recommended or consumed videos are across topics.

Table 1 presents the description and summary statistics of the three indices of content diversity, as well as other variables that we will use in the analysis.

Table 1: Description and Summary Statistics of Variables

Variables	Description	Number of Observations	Mean	St. Dev.	Min	Max
Treatment	1 if a user was assigned into the treatment group, 0	10,000,000	0.0319	0.1758	0	1
New_user	1 if a user registered on the music app after the pre-experiment period, 0	10,000,000	0.0510	0.2199	0	1
Num_registered month	Number of months a user had been registered on the app by the end of the experiment	10,000,000	40.6003	22.0691	0.0000	108.5333
Male	1 if a user was predicted to be male, 0	9,886,059	0.5344	0.4988	0	1
Age	A user's predicted age	9,734,933	23.0331	5.8964	11	45
Visit	1 if a user visited the Cloud Village in a week, 0	137,097,668	0.1972	0.3979	0	1
Freq_click	#(days in which a user watched at least one video for no less than five seconds)/7	137,097,668	0.0022	0.0244	0.0000	1.0000
Num_click	Number of clicks per week with more-than-5-second view time	137,097,668	0.2624	8.9325	0	3,655
View_min	Minutes spent watching videos per week	137,097,668	0.2132	7.0509	0.0000	4,091.2270
Num_like	Number of likes left per week	137,097,668	0.0092	0.9470	0	4,090
Num_comment	Number of comments left per week	137,097,668	0.0002	0.0404	0	167
Num_share	Number of shares left per week	137,097,668	0.0004	0.0492	0	171
Num_recommended_topic	Number of different topics recommended to a user per week when visiting the Cloud Village	27,039,411	8.1998	5.2659	1	80
HHI_recommended_topic	HHI of video topics recommended to a user per week when visiting the Cloud Village	27,039,411	0.1758	0.0741	0.0295	1.0000
Entropy_recommended_topic	Shannon entropy of video topics recommended to a user per week when visiting the Cloud Village	27,039,411	1.8912	0.4329	0.0000	3.7470
Num_clicked_topic	Number of different topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,631,675	5.5862	7.6677	1	66
HHI_clicked_topic	HHI of video topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,631,675	0.5810	0.3706	0.0391	1.0000
Entropy_clicked_topic	Shannon entropy of video topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,631,675	0.9193	0.9314	0.0000	3.4383

Notes: Summary statistics for 10 million randomly sampled Cloud Village visitors during the experiment. For 13 users receiving recommendations before they registered, we use the earliest date they received recommendations to approximate their registration date. Users' age and gender were predicted by a supervised machine learning model built by NCM. Among the sampled users, 1.14% (113,941) did not have predicted gender, and 1.83% (182,526) did not have predicted age. To remove age outliers, we keep only users in the $0.5\% - 99.5\%$ quantile of the age distribution ($11 - 45$ years old).

As a randomization check, we compare the demographics and pre-experiment recommendation diversity and activities between the treatment and control groups using regressions (see Tables 2 and 3; the regression specifications are in Online Appendix C). None of the coefficients for *Treatment* are statistically significant; thus, we conclude that there are no significant differences between the two groups.

Table 2: User Demographics at the End of the Experiment Between Control and Treatment Groups

	Dependent variable:			
	new user	*num_registered month*	*male*	*age*
Treatment	−0.00002	−0.0211	0.0001	0.0123
	(0.0004)	(0.0397)	(0.0009)	(0.0107)
Constant	0.0510***	40.6009***	0.5344***	23.0327***
	(0.0001)	(0.0071)	(0.0002)	(0.0019)
Observations	10,000,000	10,000,000	9,886,059	9,734,933
R^2	0.0000	0.000000	0.0000	0.000000

Notes: *p<0.1; **p<0.05; ***p<0.01.

Table 3: Recommendation Diversity and Users' Consumption and Engagement During the Pre-experiment Period

(a) Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.0167	−0.0002	0.0011
	(0.0136)	(0.0002)	(0.0012)
Week FE	Yes	Yes	Yes
Observations	6,386,503	6,386,503	6,386,503
R^2	0.0008	0.0015	0.0018

(b) Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.1411	−0.0035	0.0123
	(0.0928)	(0.0040)	(0.0105)
Week FE	Yes	Yes	Yes
Observations	383,371	383,371	383,371
R^2	0.0007	0.0001	0.0002

(c) Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.00001	0.00003	0.0130	0.0075	−0.0008	−0.00001	−0.00002
	(0.0004)	(0.00004)	(0.0124)	(0.0097)	(0.0005)	(0.00002)	(0.00003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	37,824,180	37,824,180	37,824,180	37,824,180	37,824,180	37,824,180	37,824,180
R^2	0.00004	0.000002	0.000001	0.000001	0.0000002	0.0000003	0.00000004

Notes: We exclude new users who registered on the music app after the pre-experiment period. Standard errors clustered at individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

In manipulation checks, we compare the recommendation diversity between the treatment and control groups during the experiment period. Results are in Table 4. In each week, treated users were recommended on average 8.316 topics compared to control users (8.196 topics), a 1.47% increase. The HHI (entropy) of recommended videos for treatment users also significantly decreased (increased) by 2.33% (1.02%), on a base of 0.176 (1.891). The changes are all statistically significant.

Table 4: Recommendation Diversity Between Treatment and Control Visitors During the Experiment

	Dependent variable:		
	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1203***	−0.0041***	0.0192***
	(0.0112)	(0.0001)	(0.0007)
Week Fixed Effects	Yes	Yes	Yes
Observations	27,039,411	27,039,411	27,039,411
R^2	0.0030	0.0097	0.0076

Notes: *p<0.1; **p<0.05; ***p<0.01.

4. Results

In this section, we first report the main effects of recommendation diversity on users' consumption behaviors, including the diversity of consumed topics and the level of consumption and engagement. We then investigate how the effects differ across new, inactive, and active users.

4.1. The Main Effects

To identify the causal effects of algorithmic recommendation diversity, we estimate the following regression model:

$$Outcome\ Variable_{it} = \beta_1 \cdot Treatment_i + u_t + e_{it} \quad (3)$$

where $Treatment_i$ is an indicator variable equal to 1 if user *i* is in the treatment group, and 0 otherwise, and u_t represents the week fixed effects. Finally, the error term e_{it} is clustered at the individual level.

For $Outcome\ Variable_{it}$, we first measure the consumption diversity using three metrics: the number ($num\_clicked\_topic_{it}$), the HHI ($HHI\_clicked\_topic_{it}$), and entropy ($entropy\_clicked\_topic_{it}$) of all video topics the user clicked on in the week.[10] Table 5(a) shows mixed regression results. While higher recommendation diversity significantly encouraged treated users to click on 0.181 more topics (3.24%, p = 0.0223), the effect based on entropy is only marginally significant (p = 0.0551), and the effect based on the HHI is insignificant (p = 0.3040). The HHI and entropy are better diversity metrics because

[10] We consider only the weeks when the user had clicked at least once in the Cloud Village; this involves 1,084,228 (10.84%) of the sampled 10 million users.

they measure how much share of each topic has changed; therefore, we conclude that there is no clear evidence suggesting that recommendation diversity helps increase users' consumption diversity.

We then investigate how the recommendation diversity affects users' consumption level, which is measured by whether the user visits the Cloud Village in a week ($visit_{it}$), the number of days in which the user *clicks* at least once ($freq\_click_{it}$), the number of *clicks* on recommended videos (num_click_{it}), and the number of minutes of watching videos ($view\_min_{it}$). We define clicks in this context as *effective* clicks, meaning that a user watched a video for at least five seconds. The measure is used by NCM to ensure that they do not count users' accidental video clicks. We also investigate the effect on the level of engagement, measured by the number of likes (num_like_{it}), comments ($num\_comment_{it}$), and shares (num_share_{it}) the user makes in the week.

Table 5(b) shows that, while there are no significant effects on most of the outcome variables, the higher recommendation diversity for treatment users significantly lowered their weekly clicking frequency by 3.08% (p = 0.0144) on a base of 0.0022. This translates to the elasticity that a 1% decrease in the recommendation HHI leads to a 1.32% drop in clicking frequency. The result is consistent with the concerns of social media platforms that more diversified recommendations could lower users' consumption level.

Table 5: The Average Treatment Effects of Recommendation Diversity

(a) Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.1809**	−0.0026	0.0138*
	(0.0791)	(0.0025)	(0.0072)
Week FE	Yes	Yes	Yes
Observations	1,631,675	1,631,675	1,631,675
R^2	0.0013	0.0008	0.0012

(b) Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0004	−0.00007**	0.0146	0.0127	0.0008	0.0001	0.00002
	(0.0003)	(0.00003)	(0.0119)	(0.0097)	(0.0010)	(0.0001)	(0.00004)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	137,097,668	137,097,668	137,097,668	137,097,668	137,097,668	137,097,668	137,097,668
R^2	0.0020	0.00005	0.00001	0.00001	0.000002	0.000001	0.000001

Notes: Standard errors clustered at individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

4.2. Heterogeneous Responses Across Users

In this subsection, we investigate whether the responses to recommendation diversity are the same across different types of users. We divide users into three groups: *new* (5.10%), *inactive* (91.42%), and *active* (3.48%). We define new users as those having not registered on the NCM app before the experiment started. They accounted for 3.42% of the total view time during the experiment. We define the remaining users as inactive (those who viewed no video during the pre-experiment period) and active (those who viewed at least one video during the period). Despite being the smallest group, active users accounted for the most view time (54.66%) during the experiment. In contrast, inactive users accounted for only 41.92% of the view time. We explore how higher recommendation diversity affects these three user groups.

4.2.1. New Users

Table 6(a) tests the manipulation effect on new users, confirming that our new algorithm significantly increased the recommendation diversity to the treated new users. Every week when they visited the Cloud Village, they were recommended 0.156 more topics ($p = 0.0037$, a 1.81% increase) than control users, facing a 2.21% ($p < 0.0001$) decrease in the HHI and a 1.02% ($p < 0.0001$) increase in entropy of recommended video topics.

Tables 6(b) and 6(c) indicate that the increase in recommendation diversity has no significant effects on the consumption diversity or the consumption and engagement levels. The results are not surprising: given that the platform has little information about new users' preferences, its algorithm cannot accurately predict the match value in the video selection stages. As such, increasing the window size for new users in our experiment should not have significant effects on their consumption and engagement behaviors.

Table 6: The Manipulation Check and Treatment Effects of Recommendation Diversity for New Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1563***	−0.0038***	0.0195***
	(0.0539)	(0.0005)	(0.0034)
Week Fixed Effects	Yes	Yes	Yes
Observations	976,454	976,454	976,454
R^2	0.0024	0.0135	0.0074

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.2879	−0.0108	0.0297
	(0.2979)	(0.0101)	(0.0277)
Week FE	Yes	Yes	Yes
Observations	69,231	69,231	69,231
R^2	0.0021	0.0021	0.0025

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0012	0.0001	0.0496	0.0513	0.0228	0.0011	0.0004
	(0.0016)	(0.0001)	(0.0588)	(0.0553)	(0.0165)	(0.0013)	(0.0002)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,230,892	4,230,892	4,230,892	4,230,892	4,230,892	4,230,892	4,230,892
R^2	0.0288	0.0008	0.0001	0.00005	0.00001	0.00002	0.00002

Notes: Standard errors clustered at individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

4.2.2. Inactive Users

For inactive users, Table 7(a) confirms that our treatment effectively increased recommendation diversity.[11] Table 7(b) shows that these users did not change their consumption diversity after being exposed to more diversified recommendations. In contrast, Table 7(c) shows that the treated users significantly reduced their clicks of videos by 5.59% (on a base of 0.0015, $p < 0.0001$). Combined with the results in Table 7(a), we find that a 1% decrease in recommendation HHI translates to a 2.60% drop in the clicking frequency.[12] The result implies that recommendation accuracy is crucial for keeping inactive users engaged. Inactive users might value the videos on the platform less, which explains why they have not watched videos in the Cloud Village for a long time. Consequently, a more diversified but less accurate recommender system could annoy them and reduce their consumption level.

[11] In Online Appendix D, we also find that the manipulation effects increased monotonically over time, which could imply the self-reinforcement of a diversity-enhanced recommender system.

[12] We also examined how higher recommendation diversity affected inactive users over time. The results show that the negative impacts on inactive users' clicking frequency first increased and then decreased as the experiment proceeded (see more detailed analysis in Online Appendix D).

Table 7: The Manipulation Check and Treatment Effects of Recommendation Diversity for Inactive Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1022***	−0.0038***	0.0177***
	(0.0082)	(0.0001)	(0.0007)
Week Fixed Effects	Yes	Yes	Yes
Observations	24,458,931	24,458,931	24,458,931
R^2	0.0038	0.0101	0.0081

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	−0.0034	0.0035	−0.0037
	(0.0553)	(0.0024)	(0.0063)
Week FE	Yes	Yes	Yes
Observations	1,154,409	1,154,409	1,154,409
R^2	0.0022	0.0018	0.0021

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0003	−0.00008***	−0.0052	−0.0016	−0.0003	0.00003	0.00003
	(0.0003)	(0.00001)	(0.0051)	(0.0048)	(0.0005)	(0.00003)	(0.00003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	127,992,368	127,992,368	127,992,368	127,992,368	127,992,368	127,992,368	127,992,368
R^2	0.0018	0.0001	0.00002	0.00002	0.000003	0.000001	0.000001

Notes: Standard errors clustered at individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

4.2.3. Active Users

Unlike new and inactive users, active users reacted positively to our new, more diversified recommender system. Table 8(a) shows that the experiment significantly decreased the HHI of recommendations[13] by 5.17% ($p < 0.0001$) for treated users on a base of 0.161. Regarding their responses, Table 8(b) shows that the HHI of clicked topics decreased by 2.85% from a base of 0.415 ($p = 0.0100$). Combined with the result in Table 8(a), the corresponding consumption diversity elasticity in response to the change in recommendation diversity is 0.55 in HHI.[14] The significant change in consumption HHI suggests that active users not only expand their topic variety but also balance more between their familiar and unfamiliar topics when facing more diversified recommendations. In contrast, we find no significant changes in active users' consumption and engagement levels. Overall, the results imply that strengthening recommendation

[13] The number of recommended topics increased by 0.330 ($p = 0.0047$), equivalent to a 2.64% increase. The entropy of recommended topics increased by 1.94% ($p < 0.0001$), on a base of 2.084.

[14] The elasticity for the number of clicked topics is 2.10 and the elasticity for entropy is 1.56. We also find that treated active users monotonically increased their consumption diversity over time (see Online Appendix D), which implies that a more diversified recommender system could induce a positive long-term effect for active users.

diversity helps increase active users' consumption diversity without hurting their important performance metrics that social media platforms are concerned about.

Table 8: The Manipulation Check and Treatment Effects of Recommendation Diversity for Active Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.3295***	−0.0083***	0.0405***
	(0.1167)	(0.0005)	(0.0044)
Week Fixed Effects	Yes	Yes	Yes
Observations	1,604,026	1,604,026	1,604,026
R^2	0.0023	0.0110	0.0098

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.5263**	−0.0118**	0.0422***
	(0.2192)	(0.0046)	(0.0154)
Week FE	Yes	Yes	Yes
Observations	408,035	408,035	408,035
R^2	0.0019	0.0010	0.0015

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0003	−0.00001	0.4591	0.3203	0.0088	−0.0004	−0.0004
	(0.0022)	(0.0007)	(0.2976)	(0.2327)	(0.0206)	(0.0005)	(0.0006)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,874,408	4,874,408	4,874,408	4,874,408	4,874,408	4,874,408	4,874,408
R^2	0.0075	0.0009	0.0003	0.0003	0.00004	0.00001	0.00003

Notes: Standard errors clustered at individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

4.3. NCM's Follow-up Experiment

Inspired by our results, NCM followed up by running a similar internal field experiment that lasted two weeks. The modified recommender system increased the window size to 15 for active users and reduced it to 1 for the others (see Online Appendix E for more details). We find that the new diversified recommendations encouraged active users to consume 3.87% more topics per week (p = 0.0002) and increased their consumption diversity in the HHI by 2.42% (p = 0.0002). Interestingly, we also find that active users' number of videos clicked and viewing time also significantly increased.[15]

Encouraged by these results, NCM permanently modified its recommender system based on our suggestions. The window size for active users has increased to 15 (from the original five). Though we

[15] Table 8(c) also shows similar positive effects on the number of videos clicked and total viewing time, though the effects are not statistically significant. The stronger effects on the consumption level in the new experiment is probably due to the difference in the window-size manipulation of the two experiments. NCM believed the window size in our experiment was too large, and they chose a smaller one.

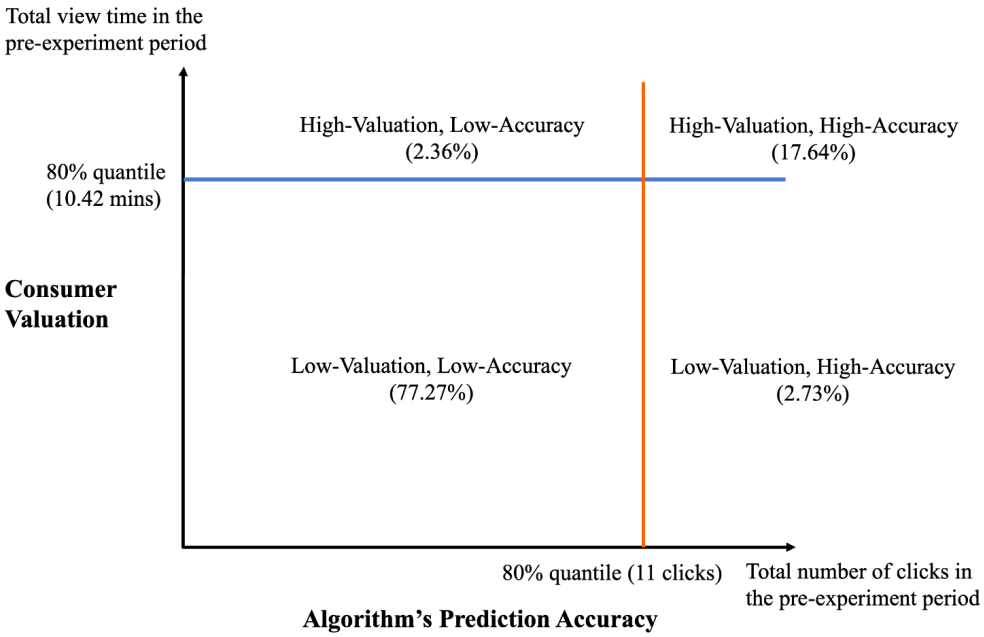
cannot determine the causal effects of this change—since there are no randomized treatment and control groups—the management revealed to us that they have seen positive responses from active users after the implementation, just as our field experiment showed.

5. Possible Mechanisms

There could be two underlying mechanisms that cause more recommendation diversity leading to more consumption diversity for active users. One, compared with other users, active users, who visit the village more often, tend to have a higher valuation of watching videos and hence are more tolerant to videos with diversified topics. This reason can also explain why inactive users' consumption diversity is not affected by recommendation diversity. Two, since active users have watched more videos in the past, the platform has collected much more information on their viewing preferences. As such, the algorithm can better predict the match value and, therefore, the modified recommender system can more successfully recommend other topics they would like to watch. This explanation is consistent with the result that new users in our treatment group did not increase their consumption diversity.

To test these two mechanisms, we divide the active users into four segments. Considering the small sample size of active users (just 348,172 of the 10 million users in our sample), we randomly resampled 2 million active users who visited the Cloud Village during the experiment for our analysis. We use the total view time during the pre-experiment period to measure active users' valuation of the Cloud Village, and we apply the 80-20 rule. Users who spent at least 10.42 minutes (80% quantile among active users) watching videos are defined as *high-valuation* users, and the others are *low-valuation* users. To measure how accurate the algorithm can predict the match value of videos for a user, we use users' total number of clicks in the pre-experiment period. As a user clicks more videos, the algorithm gathers more information about what videos the user likes and learns more about her preferences. Therefore, we classify users with at least 11 clicks (80% quantile among active users) in the pre-experiment period as *high-accuracy* users

and the others as *low-accuracy* users.[16] Figure 3 shows the four user segments: *high-valuation, high-accuracy* (17.64%), *high-valuation, low-accuracy* (2.36%), *low-valuation, high-accuracy* (2.73%), and *low-valuation, low-accuracy* (77.27%).



Notes: We randomly resample 2 million active users to guarantee the analysis's statistical power. The percentage of high-accuracy users are not exactly 20% since the click number is discrete.

Figure 3. Segmenting Active Users Based on Consumer Valuation and the Algorithm's Prediction Accuracy

Table 9 summarizes the separate estimation results for the four user segments.[17] Facing more diversified recommendations, Table 9(b) shows that high-valuation, high-accuracy users significantly increased their consumption diversity. For example, a 1% decrease in recommendation HHI led to a 0.29% ($p < 0.0001$) decrease in consumption HHI. Low-valuation, high-accuracy users also marginally reduced their consumption HHI and increased their consumption entropy. There are no effects on the consumption

[16] We also consider using the total number of impressions a user had received in the pre-experiment period to measure the algorithm's prediction accuracy. In Online Appendix F, we use both click number and impression number during the pre-experiment period to predict a user's click-through rate (i.e., the number of clicks divided by the number of impressions) in the first week of the experiment. We find that the click number has a stronger predicting power in terms of R square (0.1390 vs. 0.0164). Besides, based on feedback from NCM's data analysts, number of impressions is a noisy information about the user preference because (1) users may unintentionally click into the Cloud Village, and (2) they could be exposed to a lot of recommended videos in the Village even though they are not interested in the videos. Therefore, we choose the click number as the proxy for prediction accuracy.

[17] Online Appendix G lists more detailed regression results for the four segments.

diversity of low-accuracy users. The results suggest that the prediction accuracy of the recommender system is key for increasing the consumption diversity of users.

In contrast, Table 9(c) shows that the increased recommendation diversity significantly reduced the clicking frequency for low-valuation, low-accuracy users. Moreover, higher recommendation diversity reduced the clicking frequency, total viewing time, and the number of likes and comments of high-valuation, low-accuracy users. The results imply that, if the algorithm cannot correctly predict their video preferences, high-valuation users would not only ignore recommendations on diversified topics but also significantly reduce their consumption and engagement levels.[18] Interestingly, although low-valuation, high-accuracy users in the treatment group consumed more diversified videos, they also significantly reduced the visit probability and marginally reduced the number of comments they left. For high-valuation, high-accuracy users only, the effects on the consumption and engagement levels are not significant.[19] Overall, the results show that both high valuation of videos by users and high prediction accuracy from the algorithm are required to avoid the negative impacts of increased recommendation diversity on users' consumption and engagement levels.

[18] The effects are less negative for low-valuation, low-accuracy users probably because they have a much lower starting point.

[19] The effect on the number of shares is marginally negative. This is probably because, although these users are likely to watch the diversified topics recommended by the platform, they will not share knowing that their peers who have the same tastes may not appreciate these topics.

Table 9: The Manipulation Check and Treatment Effects of Recommendation Diversity for Four Segments of Active Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
High-Valuation, High-Accuracy User	0.6180***	−0.0134***	0.0690***
	(0.1387)	(0.0004)	(0.0043)
High-Valuation, Low-Accuracy User	−0.0216	−0.0102***	0.0410***
	(0.1999)	(0.0014)	(0.0102)
Low-Valuation, High-Accuracy User	0.4985**	−0.0092***	0.0485***
	(0.2425)	(0.0013)	(0.0106)
Low-Valuation, Low-Accuracy User	0.2042***	−0.0069***	0.0322***
	(0.0306)	(0.0002)	(0.0017)

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
High-Valuation, High-Accuracy User	0.3830***	−0.0080***	0.0342***
	(0.1313)	(0.0020)	(0.0076)
High-Valuation, Low-Accuracy User	−0.2510	0.0174	−0.0407
	(0.2767)	(0.0110)	(0.0313)
Low-Valuation, High-Accuracy User	0.4706	−0.0146*	0.0463*
	(0.3030)	(0.0081)	(0.0256)
Low-Valuation, Low-Accuracy User	0.0779	0.0022	0.0015
	(0.0714)	(0.0028)	(0.0077)

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
High-Valuation, High-Accuracy User	0.0013	0.0008	0.7386	0.6262	0.0178	−0.0014	−0.0027*
	(0.0027)	(0.0013)	(0.5690)	(0.4594)	(0.0434)	(0.0015)	(0.0014)
High-Valuation, Low-Accuracy User	−0.0002	−0.0022***	−0.1894	−0.3205**	−0.0261***	−0.0015***	0.0018
	(0.0059)	(0.0008)	(0.1840)	(0.1570)	(0.0046)	(0.0003)	(0.0018)
Low-Valuation, High-Accuracy User	−0.0163***	−0.0002	0.4060	0.3240	−0.0189	−0.0010*	−0.0008
	(0.0056)	(0.0012)	(0.4118)	(0.3256)	(0.0189)	(0.0006)	(0.0011)
Low-Valuation, Low-Accuracy User	0.0003	−0.0002**	0.0217	0.0151	−0.0022	0.0002	−0.00002
	(0.0010)	(0.0001)	(0.0323)	(0.0251)	(0.0018)	(0.0002)	(0.0001)

Notes: Estimates for the difference between treatment and control groups after controlling for week fixed effects. There are 352,756 high-valuation, high-accuracy users, 47,246 high-valuation, low-accuracy users, 54,574 low-valuation, high-accuracy users, and 1,545,424 low-valuation, low-accuracy users. Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

There could be other alternative mechanisms. For example, as documented by Hu and Pu (2011), users' perceived diversity may differ from the algorithm's designed diversity level. Compared to active users, new and inactive users may be less sensitive to the increase in recommendation diversity. However, given that our algorithm has significantly increased the diversity of recommendations for these users, they should have consumed more diversely if they did not perceive the algorithm change. Since neither new nor inactive users changed their consumption diversity, we can rule out this alternative explanation.

6. Conclusion

We conducted a field experiment with NCM to evaluate the effects of content diversity in algorithmic recommendations on social media users' consumption diversity, as well as on their consumption and

engagement levels. Our results show that the increased recommendation diversity, on average, did not have a strong effect on consumption diversity but reduced consumption level. The effects, however, differ across new, inactive, and active users. New users were not affected by the increased recommendation diversity, and inactive users reduced only clicks on videos without changing their consumption diversity or engagement level. In contrast, active users, who contribute most to the platform's viewing time, not only maintained their consumption and engagement levels but also increased their consumption diversity when facing more diversified recommendations. We further explore the possible mechanisms for the positive responses from active users. We find that they increase consumption diversity only if their preferences are well understood by the platform, and that their consumption level will not be hurt if they also highly value the platform's videos.

Our findings suggest a recommendation strategy for social media platforms like NCM. For new and inactive users, platforms should set a lower level of diversity to help ensure recommendation accuracy. By enticing these users visit and consume content, the recommender system can learn their preferences over time. After these users grow into high-valuation users and the recommender algorithm has well learned their preferences, platforms can increase the diversity of recommendations. This way, while helping users get out of their social media bubbles, platforms can also explore users' interests, which will improve long-term profits.

Our research also has important implications for policymakers. First, we demonstrate that it is possible for platforms to push for more diverse consumption of digital content. Second, we shed light on the intersection between public policy about user data privacy and diversified consumption. To encourage users to consume more diversified content on digital platforms, platforms need to understand users' consumption preferences; otherwise users would ignore the broader recommended contents and stay in their social media bubbles. Therefore, limiting digital platforms' use of individual data might block a way for users to explore new content and embrace more diversity. This insight points to a complicated trade-off for regulators who seek to block platforms from using individual users' browsing and clicking data.

Our paper has several limitations that should be addressed by future research. First, we conducted our experiment on a video-based social media platform. It would be interesting to extend our results to other text-based or photo-based social media platforms. Second, our experiment was built directly on the platform's cutting-edge recommender system, which could underestimate the treatment effect due to spillovers. Specifically, when we increased the recommendation diversity for treated users, the recommended videos would get more impressions and usually receive more clicks and likes. This process can affect the control group, since the algorithm may increase the recommendation of these videos. Because only 3% of users were assigned to the treatment group, we believe the spillover is limited. Investigating the dynamic spillover effects on algorithm recommendations when a platform widely adopts the increase in recommendation diversity is important for future researchers. Third, while we focus on one particular way of improving recommendation diversity, it would also be interesting to study how other ways of modifying existing recommender systems to increase diversity could affect users' behavior. Finally, consumer valuation and the algorithm's prediction accuracy are only two possible mechanisms for explaining the heterogeneous responses among users. Other potential mechanisms should also be explored in the future.

References

Adomavicius G, Kwon Y (2011a) Maximizing aggregate recommendation diversity: A graph-theoretic approach. *Proc. 1st Int. Work. Nov. Divers. Recomm. Syst. (DiveRS 2011)*. (Citeseer), 3–10.

Adomavicius G, Kwon Y (2011b) Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Trans. Knowl. Data Eng.* 24(5):896–911.

Anderson A, Maystre L, Anderson I, Mehrotra R, Lalmas M (2020) Algorithmic Effects on the Diversity of Consumption on Spotify. *Web Conf. 2020 - Proc. World Wide Web Conf. WWW 2020* 2:2155–2165.

Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. *Science (80-.).* 348(6239):1130–1132.

Berman R, Katona Z (2020) Curation algorithms and filter bubbles in social networks. *Mark. Sci.* 39(2):296–316.

Bradley K, Smyth B (2001) Improving Recommendation Diversity. *Business*:75–84.

Clarke CLA, Kolla M, Cormack G V., Vechtomova O, Ashkan A, Büttcher S, MacKinnon I (2008) Novelty and diversity in information retrieval evaluation. *ACM SIGIR 2008 - 31st Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, Proc.*:659–666.

Claussen J, Peukert C, Sen A (2019) The Editor vs. The Algorithm: Returns to Data and Externalities in Online News. *SSRN Electron. J.*

Dujeancourt E, Garz M, Ghose A, Hagen J, Lischka J (2022) The effects of algorithmic content selection on user engagement with news on twitter.

Ekstrand MD, Harper FM, Willemsen MC, Konstan JA (2014) User perception of differences in recommender algorithms. *RecSys 2014 - Proc. 8th ACM Conf. Recomm. Syst.*:161–168.

Fleder D, Hosanagar K (2009) Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Manage. Sci.* 55(5):697–712.

Ghose A, Ipeirotis PG, Li B (2014) Examining the impact of ranking on consumer behavior and search engine revenue. *Manage. Sci.* 60(7):1632–1654.

Holtz D, Carterette B, Chandar P, Nazari Z, Cramer H, Aral S (2020) The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify. *EC 2020 - Proc. 21st ACM Conf. Econ. Comput.*:75–76.

Hosanagar K, Fleder D, Lee D, Buja A (2014) Will the global village fracture into tribes recommender systems and their effects on consumer fragmentation. *Manage. Sci.* 60(4):805–823.

Hu R, Pu P (2011) Helping Users Perceive Recommendation Diversity. *Divers. RecSys*. 43–50.

Hurley N, Zhang M (2011) Novelty and Diversity in Top-N Recommendation -- Analysis and Evaluation. *ACM Trans. Internet Technol.* 10(4):30.

Huszár F, Ktena SI, OBrien C, Belli L, Schlaikjer A, Hardt M (2022) Algorithmic amplification of politics on Twitter. *Proc. Natl. Acad. Sci. U. S. A.* 119(1):1–6.

Javari A, Jalili M (2015) A probabilistic model to resolve diversity–accuracy challenge of recommendation systems. *Knowl. Inf. Syst.* 44(3):609–627.

Kaminskas M, Bridge D (2016) Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-Accuracy objectives in recommender systems. *ACM Trans. Interact. Intell. Syst.* 7(1):1–42.

Kunaver M, Požrl T (2017) Diversity in recommender systems – A survey. *Knowledge-Based Syst.* 123:154–162.

Lee D, Hosanagar K (2019) How do recommender systems affect sales diversity? A cross-category investigation via

randomized field experiment. *Inf. Syst. Res.* 30(1):239–259.

Moehring A (2022) *News Feeds and User Engagement: Evidence from the Reddit News Tab*

Pariser E (2011) *The filter bubble: How the new personalized web is changing what we read and how we think*

Ribeiro MH, Ottoni R, West R, Almeida VAF, Wagner Meira WM (2020) Auditing radicalization pathways on YouTube. *FAT* 2020 - Proc. 2020 Conf. Fairness, Accountability, Transpar.*:131–141.

Song M (2021) *How Do Personalized Recommendations Affect Consumer Exploration: A Field Experiment*

Vaishnavi S, Jayanthi A, Karthik S (2013) Ranking technique to improve diversity in recommender systems. *Int. J. Comput. Appl.* 68(2).

Van Dam A (2019) Diversity and its decomposition into variety, balance and disparity. *R. Soc. Open Sci.* 6(7).

Vargas S, Baltrunas L, Karatzoglou A, Castells P (2014) Coverage, redundancy and size-awareness in genre diversity for recommender systems. *RecSys 2014 - Proc. 8th ACM Conf. Recomm. Syst.*:209–216.

Vargas S, Castells P (2011) Rank and relevance in novelty and diversity metrics for recommender systems. *RecSys'11 - Proc. 5th ACM Conf. Recomm. Syst.*:109–116.

Zhou R, Khemmarat S, Gao L (2010) The impact of YouTube recommendation system on video views. *Proc. ACM SIGCOMM Internet Meas. Conf. IMC*:404–410.

Ziegler CN, McNee SM, Nr G, Konstan JA, Lausen G (2005) Improving recommendation lists through topic diversification. :22.

Online Appendix

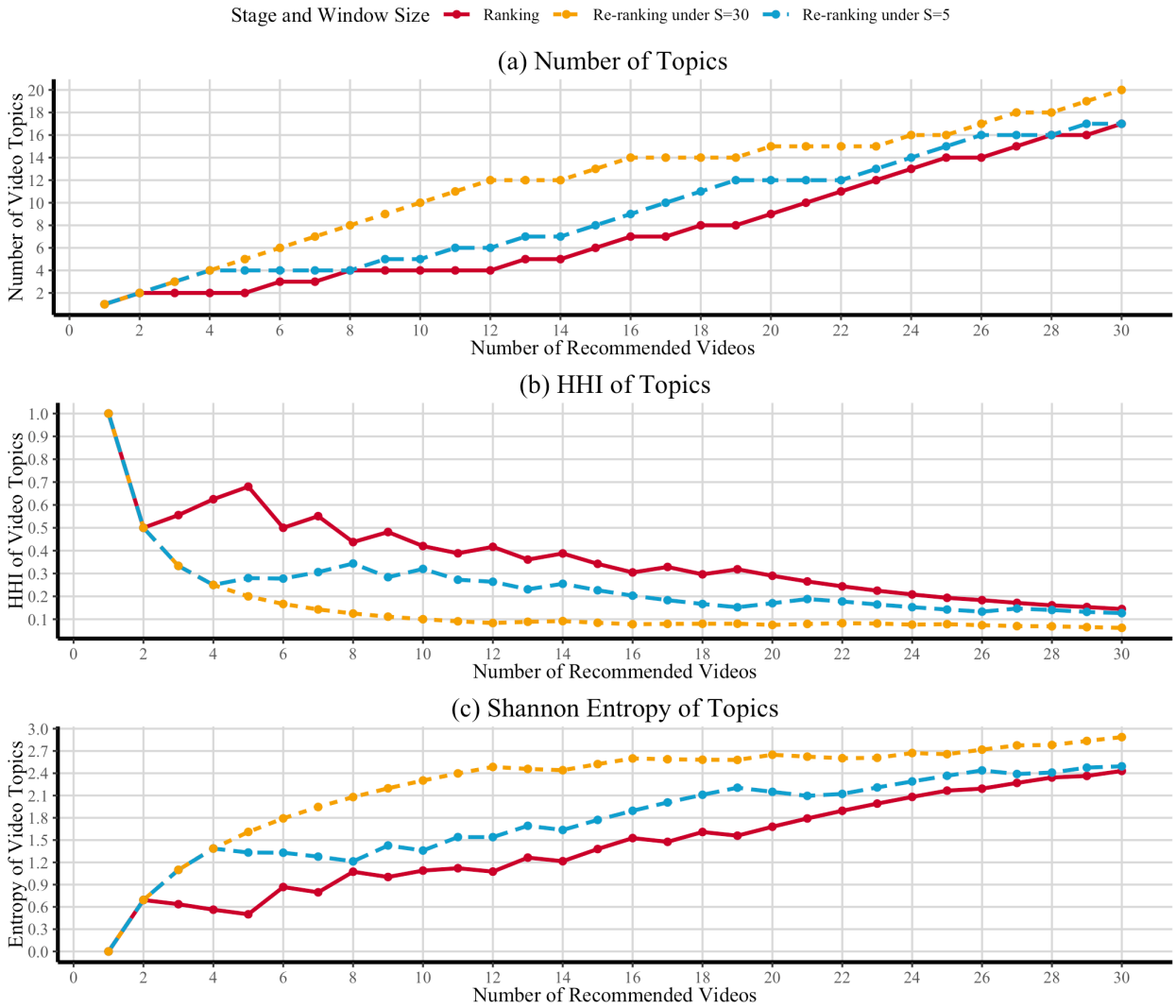
Appendix A: An Example of Algorithmic Recommendations Under Different Window Sizes

We randomly select a Cloud Village user and simulate the algorithm's recommendations for her one query. To illustrate why increasing window size can effectively increase the content diversity of recommendations, we ask the recommender algorithm[20] to retrieve 400 videos, select 50 top-ranked videos based on their predicted match value to the user, and re-rank the 50 videos using window size 5 or 30.

Table A1 shows the relationship between the number of recommended videos and the topic diversity of recommendations in the ranking stage and re-ranking stage with different window sizes.[21] First, in the ranking stage (red line), as the algorithm recommends more videos to the user, the diversity of recommendations overall increases in terms of topic number, the HHI of topics, and Shannon entropy of topics. It implies that the recommender system can retrieve more diverse and unfamiliar video topics for the user, despite giving these videos a lower ranking. Therefore, when we expand the window size, the re-ranking algorithm is more likely to include these more diverse videos in the consideration set and recommend them to the user. To further prove this point, we compare the algorithm's final recommendations after it re-ranks videos using the window size of 5 (blue line) and 30 (orange line). Conditional on recommending the same number of videos, the algorithm under window size 30 indeed selects more diverse videos.[22] For example, suppose the algorithm recommends 10 videos to a user per request. Under window size 5, the algorithm recommends five different topics with an HHI of 0.32 and Shannon entropy of 1.36. By contrast, under window size 30, the algorithm recommends 10 different topics with an HHI of 0.10 and Shannon entropy of 2.30. Therefore, we confirm that a larger window size effectively increases the content diversity of recommendations in the Cloud Village.

[20] For simplicity, here we incorporate the re-prediction of user-video match values in the Ranking stage and emphasize only the window-size change in the Re-ranking stage.
[21] We assume the algorithm recommends at most 30 videos per user query, which is consistent with the setting in NCM's recommender system.
[22] In this example, the recommendation diversity under two window sizes is the same when the algorithm recommends four or fewer videos to the user. In reality, the recommender algorithm recommends far more than four videos per user query.

Table A1: The Diversity of Video Topics Under Different Recommendation Stages and Window Sizes



Appendix B: Treatment Effect Estimation by Including the Experiment Implementation Period in the Experiment Period

We show that the estimation effects are qualitatively consistent if we include the experiment implementation period in the experiment period. Specifically, we split the data into two periods: (1) the pre-experiment period, which comprises the four weeks before the experiment started (November 19 to December 16, 2021), and (2) the experiment period, totaling 14 weeks after December 17, 2021 (December 18, 2021, to March 25, 2022). Among the sampled 10 million users, we filter 9,821,061 users who had visited the Cloud Village during the experiment period, including 485,152 (4.94%) new users, 8,989,773 (91.54%) inactive users, and 346,136 (3.52%) active users.[23] Table B1 shows the description and summary statistics of regression variables.

[23] In total, 9,821,066 users visited the Cloud Village during the experiment period. We removed five users who had received recommendations before the registration date.

Table B1: Description and Summary Statistics of Variables

Variables	Description	Number of Observations	Mean	St. Dev.	Min	Max
Treatment	1 if a user was assigned into the treatment group, 0	9,821,061	0.0319	0.1758	0	1
New_user	1 if a user registered on the music app after the pre-experiment period, 0	9,821,061	0.0494	0.2167	0	1
Num_registered month	Number of months a user had been registered on the app by the end of the experiment	9,821,061	40.5598	22.0299	0.0000	108.4000
Male	1 if a user was predicted to be male, 0	9,821,061	0.5345	0.4988	0	1
Age	A user's predicted age	9,577,475	23.0285	5.8925	11	45
Visit	1 if a user visited the Cloud Village in a week, 0	134,605,452	0.1994	0.3996	0	1
Freq_click	#(days in which a user watched at least one video for no less than five seconds)/7	134,605,452	0.0022	0.0246	0.0000	1.0000
Num_click	Number of clicks per week with more-than-5-second view time	134,605,452	0.2665	9.0086	0	4,219
View_min	Minutes spent watching videos per week	134,605,452	0.2165	7.0974	0.0000	4,730.8230
Num_like	Number of likes left per week	134,605,452	0.0093	0.9563	0	4,267
Num_comment	Number of comments left per week	134,605,452	0.0002	0.0403	0	142
Num_share	Number of shares left per week	134,605,452	0.0004	0.0492	0	195
Num_recommended_topic	Number of different topics recommended to a user per week when visiting the Cloud Village	26,846,219	8.2227	5.2693	1	79
HHI_recommended_topic	HHI of video topics recommended to a user per week when visiting the Cloud Village	26,846,219	0.1758	0.0749	0.0302	1.0000
Entropy_recommended_topic	Shannon entropy of video topics recommended to a user per week when visiting the Cloud Village	26,846,219	1.8927	0.4352	0.0000	3.7212
Num_clicked_topic	Number of different topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,617,598	5.6072	7.6738	1	65
HHI_clicked_topic	HHI of video topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,617,598	0.5797	0.3708	0.0386	1.0000
Entropy_clicked_topic	Shannon entropy of video topics clicked (i.e., viewed for at least five seconds) by a user per week when having clicks	1,617,598	0.9227	0.9327	0.0000	3.4800

Notes: Summary statistics for 9,821,061 Cloud Village visitors during the experiment. Users' age and gender were predicted by a supervised machine learning model built by NCM. Among the sampled users, 1.03% (101,638) did not have predicted gender, and 1.65% (162,326) did not have predicted age. To remove age outliers, we keep only users in the $0.5\% - 99.5\%$ quantile of the age distribution ($11 - 45$ years old).

As Tables B2 and B3 show, the randomization check also passes, since there is no significant difference between treatment and control users in their demographics, pre-experiment recommendation diversity, or pre-experiment activities.

Table B2: User Demographics at the End of the Experiment Between Control and Treatment Groups

	Dependent variable:			
	new user	*num_registered month*	*male*	*age*
Treatment	0.00001	−0.0195	0.0003	0.0138
	(0.0004)	(0.0400)	(0.0009)	(0.0108)
Constant	0.0494***	40.5604***	0.5345***	23.0281***
	(0.0001)	(0.0071)	(0.0002)	(0.0019)
Observations	9,821,061	9,821,061	9,719,423	9,577,475
R^2	0.0000	0.000000	0.0000	0.000000

Notes: *p<0.1; **p<0.05; ***p<0.01.

Table B3: Recommendation Diversity and Users' Consumption and Engagement During the Pre-experiment Period

(a) Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.0167	−0.0002	0.0012
	(0.0137)	(0.0002)	(0.0012)
Week FE	Yes	Yes	Yes
Observations	6,343,722	6,343,722	6,343,722
R^2	0.0008	0.0015	0.0018

(b) Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.1370	−0.0031	0.0115
	(0.0932)	(0.0040)	(0.0106)
Week FE	Yes	Yes	Yes
Observations	381,491	381,491	381,491
R^2	0.0007	0.0001	0.0002

(c) Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.0001	0.00002	0.0129	0.0074	−0.0008	−0.00001	−0.00002
	(0.0005)	(0.00004)	(0.0126)	(0.0099)	(0.0005)	(0.00002)	(0.00004)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	37,207,613	37,207,613	37,207,613	37,207,613	37,207,613	37,207,613	37,207,613
R^2	0.00004	0.000002	0.000001	0.000001	0.0000002	0.0000003	0.00000004

Notes: We exclude new users who registered on the music app after the pre-experiment period. Standard errors clustered at the individual level are in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table B4 shows the main effects of recommendation diversity on users' consumption behavior. Consistent with the results in Section 4.1, despite 2.27% ($p < 0.0001$) increased recommendation diversity in the HHI, we find no clear evidence suggesting that our new algorithm increased users' consumption diversity. While higher recommendation diversity significantly encouraged treated users to click on 0.177 more topics (3.16%, $p = 0.0260$), the effect based on entropy is only marginally significant ($p = 0.0617$), and the effect based on the HHI is insignificant ($p = 0.3320$). However, higher recommendation diversity significantly reduced users' clicking frequency by 2.99% ($p = 0.0184$).

Table B4: The Manipulation Check and Average Treatment Effects of Recommendation Diversity

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1191***	−0.0040***	0.0190***
	(0.0113)	(0.0001)	(0.0007)
Week Fixed Effects	Yes	Yes	Yes
Observations	26,846,219	26,846,219	26,846,219
R^2	0.0019	0.0054	0.0044

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.1770**	−0.0024	0.0136*
	(0.0795)	(0.0025)	(0.0073)
Week FE	Yes	Yes	Yes
Observations	1,617,598	1,617,598	1,617,598
R^2	0.0014	0.0010	0.0014

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0003	−0.00007**	0.0142	0.0119	0.0008	0.0001	0.00002
	(0.0003)	(0.00003)	(0.0120)	(0.0097)	(0.0010)	(0.0001)	(0.00003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	134,605,452	134,605,452	134,605,452	134,605,452	134,605,452	134,605,452	134,605,452
R^2	0.0029	0.00005	0.00001	0.00001	0.000002	0.000001	0.000001

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Regarding the heterogeneous responses across users, Table B5 confirms that more recommendation diversity did not affect new users' consumption diversity or consumption and engagement levels. Table B6 shows that, facing 2.14% increased recommendation diversity in the HHI ($p < 0.0001$), inactive users significantly reduced their weekly clicking frequency by 5.51% ($p < 0.0001$) without consuming more diversely. A 1% decrease in the recommendation HHI translates into a 2.57% drop in inactive users' clicking frequency. Table B7 also confirms that more recommendation diversity increased active users' consumption diversity without hurting their level of consumption or engagement. A 5.00% ($p < 0.0001$) decrease in the recommendation HHI translated into a 2.30% ($p = 0.0355$) decrease in their consumption HHI. Therefore, all results are qualitatively consistent with Section 4.2.

Table B5: The Manipulation Check and Treatment Effects of Recommendation Diversity for New Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1643***	−0.0036***	0.0191***
	(0.0547)	(0.0005)	(0.0035)
Week Fixed Effects	Yes	Yes	Yes
Observations	926,100	926,100	926,100
R^2	0.0018	0.0106	0.0054

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.3152	−0.0089	0.0285
	(0.3026)	(0.0103)	(0.0285)
Week FE	Yes	Yes	Yes
Observations	65,478	65,478	65,478
R^2	0.0027	0.0026	0.0030

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0005	0.00004	0.0444	0.0477	0.0227	0.0011	0.0004
	(0.0016)	(0.0001)	(0.0593)	(0.0543)	(0.0169)	(0.0013)	(0.0003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,902,726	3,902,726	3,902,726	3,902,726	3,902,726	3,902,726	3,902,726
R^2	0.0313	0.0009	0.0001	0.00005	0.00001	0.00002	0.00001

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table B6: The Manipulation Check and Treatment Effects of Recommendation Diversity for Inactive Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1011***	−0.0038***	0.0174***
	(0.0082)	(0.0001)	(0.0007)
Week Fixed Effects	Yes	Yes	Yes
Observations	24,296,402	24,296,402	24,296,402
R^2	0.0024	0.0057	0.0047

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	−0.0010	0.0030	−0.0024
	(0.0547)	(0.0024)	(0.0063)
Week FE	Yes	Yes	Yes
Observations	1,136,224	1,136,224	1,136,224
R^2	0.0027	0.0022	0.0025

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0003	−0.00008***	−0.0056	−0.0020	−0.0003	0.00004	0.00002
	(0.0003)	(0.00001)	(0.0050)	(0.0047)	(0.0005)	(0.00003)	(0.00003)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	125,856,822	125,856,822	125,856,822	125,856,822	125,856,822	125,856,822	125,856,822
R^2	0.0029	0.0002	0.00003	0.00003	0.000004	0.000002	0.000001

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table B7: The Manipulation Check and Treatment Effects of Recommendation Diversity for Active Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.3258***	−0.0080***	0.0398***
	(0.1154)	(0.0005)	(0.0044)
Week Fixed Effects	Yes	Yes	Yes
Observations	1,623,717	1,623,717	1,623,717
R^2	0.0020	0.0079	0.0078

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.4801**	−0.0095**	0.0361**
	(0.2164)	(0.0045)	(0.0153)
Week FE	Yes	Yes	Yes
Observations	415,896	415,896	415,896
R^2	0.0021	0.0011	0.0017

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0001	0.00002	0.4649	0.3124	0.0090	−0.0004	−0.0004
	(0.0022)	(0.0007)	(0.2984)	(0.2324)	(0.0206)	(0.0005)	(0.0006)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,845,904	4,845,904	4,845,904	4,845,904	4,845,904	4,845,904	4,845,904
R^2	0.0083	0.0011	0.0003	0.0003	0.00004	0.00001	0.00003

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Last, to confirm that the results of mechanism analysis are also consistent with Section 5, we keep the 1,988,668 active users[24] who had visited the Cloud Village from December 18, 2021, to March 25, 2022. Following the analysis in the main paper, we classify these active users into four segments by their 80% quantile of view time (10.46 minutes) and clicking number (11) during the pre-experiment period: *high-valuation, high-accuracy* (17.66%), *high-valuation, low-accuracy* (2.34%), *low-valuation, high-accuracy* (2.75%), and *low-valuation, low-accuracy* (77.25%).

Tables B8–B11 show that the estimation results are qualitatively consistent with the main paper. For high-valuation, high-accuracy users only, higher recommendation diversity encouraged more diverse consumption without significantly hurting important performance metrics. The 8.79% decrease in the recommendation HHI led to a 2.65% decrease in the consumption HHI, translating into an elasticity of 0.30. For low-accuracy users, more diversified recommendations did not affect consumption diversity but

[24] In another words, we remove the active users in the 2 million sample who did not visit the Cloud Village before March 26, 2022.

significantly reduced performance metrics. Facing a 1% decrease in the recommendation HHI, high-valuation, low-accuracy users clicked 2.10% less frequently, viewed 3.43% less time, and left 8.64% fewer likes and 11.01% fewer comments; low-valuation, low-accuracy users clicked 0.60% less frequently. For low-valuation, high-accuracy users, though higher recommendation diversity marginally lifted their consumption diversity with an elasticity of 0.61 in the HHI (p = 0.0595), it significantly reduced their visiting probability with an elasticity of 0.80 in the HHI (p = 0.0036).[25]

Table B8: The Manipulation Check and Treatment Effects of Recommendation Diversity for High-Valuation, High-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.6129***	−0.0129***	0.0673***
	(0.1369)	(0.0004)	(0.0042)
Week FE	Yes	Yes	Yes
Observations	2,157,064	2,157,064	2,157,064
R^2	0.0029	0.0162	0.0119

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.3859***	−0.0079***	0.0350***
	(0.1299)	(0.0019)	(0.0075)
Week FE	Yes	Yes	Yes
Observations	1,246,870	1,246,870	1,246,870
R^2	0.0027	0.0020	0.0027

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0017	0.0010	0.8081	0.6647	0.0160	−0.0014	−0.0026*
	(0.0026)	(0.0013)	(0.5758)	(0.4637)	(0.0443)	(0.0015)	(0.0015)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,916,058	4,916,058	4,916,058	4,916,058	4,916,058	4,916,058	4,916,058
R^2	0.0114	0.0052	0.0021	0.0021	0.0003	0.00003	0.0001

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

[25] The experiment significantly decreased the recommendation HHI by 5.75% for high-valuation, low-accuracy users, 4.01% for low-valuation, low-accuracy users, and 5.71% for low-valuation, high-accuracy users.

Table B9: The Manipulation Check and Treatment Effects of Recommendation Diversity for High-Valuation, Low-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	−0.0351	−0.0096***	0.0373***
	(0.1980)	(0.0013)	(0.0100)
Week FE	Yes	Yes	Yes
Observations	214,018	214,018	214,018
R^2	0.0018	0.0073	0.0072

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	−0.2063	0.0120	−0.0284
	(0.2696)	(0.0107)	(0.0305)
Week FE	Yes	Yes	Yes
Observations	57,421	57,421	57,421
R^2	0.0041	0.0014	0.0023

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0005	−0.0021***	−0.1889	−0.3183**	−0.0251***	−0.0015***	0.0017
	(0.0060)	(0.0008)	(0.1828)	(0.1548)	(0.0048)	(0.0004)	(0.0018)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	652,218	652,218	652,218	652,218	652,218	652,218	652,218
R^2	0.0078	0.0009	0.0001	0.0001	0.00004	0.00002	0.0001

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table B10: The Manipulation Check and Treatment Effects of Recommendation Diversity for Low-Valuation, High-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.4626*	−0.0089***	0.0468***
	(0.2382)	(0.0012)	(0.0103)
Week FE	Yes	Yes	Yes
Observations	272,165	272,165	272,165
R^2	0.0016	0.0069	0.0065

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.3419	−0.0147*	0.0395
	(0.2865)	(0.0078)	(0.0246)
Week FE	Yes	Yes	Yes
Observations	95,250	95,250	95,250
R^2	0.0050	0.0009	0.0020

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.0163***	−0.0003	0.3525	0.2718	−0.0196	−0.0010*	−0.0009
	(0.0056)	(0.0012)	(0.3981)	(0.3105)	(0.0186)	(0.0006)	(0.0010)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	766,108	766,108	766,108	766,108	766,108	766,108	766,108
R^2	0.0084	0.0014	0.0002	0.0002	0.0001	0.00002	0.0001

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table B11: The Manipulation Check and Treatment Effects of Recommendation Diversity for Low-Valuation, Low-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1930***	−0.0066***	0.0306***
	(0.0301)	(0.0002)	(0.0017)
Week FE	Yes	Yes	Yes
Observations	6,675,641	6,675,641	6,675,641
R^2	0.0028	0.0072	0.0083

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.0917	0.0008	0.0049
	(0.0694)	(0.0028)	(0.0076)
Week FE	Yes	Yes	Yes
Observations	991,250	991,250	991,250
R^2	0.0037	0.0016	0.0024

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0010	−0.0002**	0.0202	0.0139	−0.0019	0.0003	−0.00003
	(0.0010)	(0.0001)	(0.0313)	(0.0243)	(0.0018)	(0.0002)	(0.0001)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	21,506,968	21,506,968	21,506,968	21,506,968	21,506,968	21,506,968	21,506,968
R^2	0.0078	0.0003	0.00004	0.00003	0.00001	0.00001	0.000005

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Appendix C: Regression Specifications for Randomization Check

We compare treatment and control users' demographics using the following regression specification:

$$Demographics_i = \gamma_1 \cdot Treatment_i + \eta_i \qquad (C1)$$

where $Demographics_i \in \{new\ user_i, num\_registered\ month_i, male_i, age_i\}$ is detailed later. $Treatment_i$ is an indicator variable equal to 1 if user *i* is in the treatment group, and 0 otherwise. η_i is the individual-specific error term. We define new users as those who registered on the app after the pre-experiment period ($new\ user_i$), and we calculate each user *i*'s number of registered months till the end of the experiment ($num\_registered\ month_i$). Each user *i*'s age and gender are predicted by a well-trained, supervised, deep-learning model built by NCM. $male_i$ is 1 if the user is predicted to be male and 0 otherwise.

For users who had registered before the experiment implementation period, we also checked whether the treatment and control groups were recommended videos of different diversity levels or behaved differently during the pre-experiment period. The regression model is specified as follows:

$$Outcome\ Variable_{it} = \lambda_1 \cdot Treatment_i + \theta_t + \tau_{it}, \qquad (C2)$$

where $Outcome\ Variable_{it}$ for user *i* in week *t* is detailed later. $Treatment_i$ is a binary variable indicating whether user *i* was in the treatment (vs. control) condition. θ_t represents the week fixed effects, and the error term τ_{it} is clustered at the individual level.

The outcome variables consist of three sets of measures: the topic diversity of recommended videos when user *i* visited the Cloud Village in week *t* including the number ($num\_recommended\_topic_{it}$), the HHI ($HHI\_recommended\_topic_{it}$), and the Shannon entropy of recommended video topics ($entropy\_recommended\_topic_{it}$); user *i*'s consumption diversity when she clicked on videos in week t including the number ($num\_clicked\_topic_{it}$), the HHI ($HHI\_clicked\_topic_{it}$), and the Shannon entropy of clicked video topics ($entropy\_clicked\_topic_{it}$); user *i*'s weekly consumption and engagement levels including whether the user visited the Cloud Village ($visit_{it}$), the number of days having clicks ($freq\_click_{it}$), the number of clicks on recommended videos (num_click_{it}), the number of minutes spent

watching videos ($view\_min_{it}$), and the numbers of likes (num_click_{it}), comments ($num\_comment_{it}$), and shares (num_share_{it}) the user made.

Appendix D: Time-varying Treatment Effects for Inactive and Active Users

To examine whether the effects of recommendation diversity vary over time, we divide the 14-week experiment periods into four months and estimate the treatment effects for inactive and active users in each month.[26]

Table D1 summarizes the manipulation and treatment effects of recommendation diversity for inactive users over time. Consistent with the results in Section 4.2.2, there are no significant and clear effects on active users' consumption diversity or engagement level. However, for their consumption level measured by the clicking frequency, we find that the negative impacts of recommendation diversity first increased then decreased as the experiment proceeded (see Table D1(c), column 2). Table D2 also shows the similar time trend for the elasticity of clicking frequency. Facing a 1% decrease in the recommendation HHI, inactive users reduced their clicking frequency by 2.30% in the first month of the experiment. They continued to click even less frequently until the third month. In the last month, though the effect was still negative, inactive users only lowered their clicking frequency by 1.56%. It seems that inactive users tended to get used to the algorithm change after three months of adaptation.

For active users, as Table D3 shows, we still find no significant effects of recommendation diversity on their consumption or engagement levels. More importantly, we observe a monotonically increasing effect of recommendation diversity on their consumption diversity (see Table D3(b), column 2). In the first two months of the experiment, treated users did not significantly change their consumption diversity in the HHI; however, starting from the third month, they began to significantly and continuously increase their consumption diversity in the HHI. Table D4 further demonstrates that active users' elasticity of consumption diversity gradually increased over time, from 0.41 in the first month to 0.79 in the last month.

One more interesting phenomenon is that the manipulation effects for both inactive and active users also monotonically increased over time. For example, as shown in Tables D2(a) and D4(a), the algorithm recommended more and more diverse videos to treated users as the experiment proceeded. We conjecture

[26] Each of the first three months contains four weeks, and the last "month" contains two weeks.

that, though our experiment changed the algorithm only once, the recommender system could self-reinforce the effects and continuously improve diversification in the long run. Specifically, our diversity-enhanced algorithm for treatment users might recommend more niche videos to users and help these videos win more clicks and likes. Correspondingly, the recommender system would rank them higher in future recommendations and be more likely to recommend even more diverse videos to treatment users.

Table D1: The Manipulation Check and Treatment Effects of Recommendation Diversity for Inactive Users Over Time

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
$Treatment \times Month\ 1$	0.0942***	−0.0030***	0.0143***
	(0.0112)	(0.0002)	(0.0010)
$Treatment \times Month\ 2$	0.1068***	−0.0039***	0.0186***
	(0.0123)	(0.0002)	(0.0010)
$Treatment \times Month\ 3$	0.0930***	−0.0041***	0.0180***
	(0.0123)	(0.0002)	(0.0010)
$Treatment \times Month\ 4$	0.1290***	−0.0049***	0.0222***
	(0.0162)	(0.0002)	(0.0013)
Week Fixed Effects	Yes	Yes	Yes
Observations	24,458,931	24,458,931	24,458,931
R^2	0.0038	0.0101	0.0081

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
$Treatment \times Month\ 1$	0.1167	−0.0046	0.0158
	(0.0764)	(0.0043)	(0.0104)
$Treatment \times Month\ 2$	−0.0653	0.0053	−0.0095
	(0.0779)	(0.0038)	(0.0096)
$Treatment \times Month\ 3$	−0.0697	0.0107***	−0.0177*
	(0.0815)	(0.0040)	(0.0100)
$Treatment \times Month\ 4$	0.0527	−0.0004	0.0027
	(0.1077)	(0.0053)	(0.0131)
Week Fixed Effects	Yes	Yes	Yes
Observations	1,154,409	1,154,409	1,154,409
R^2	0.0022	0.0018	0.0021

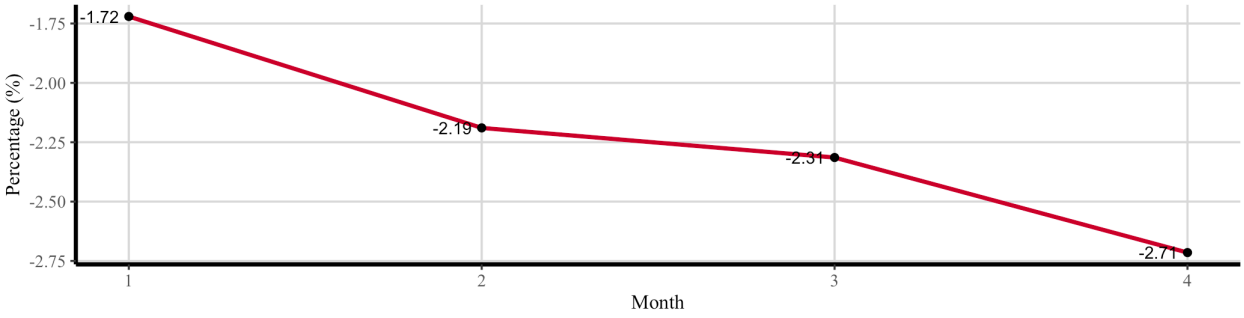
(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
$Treatment \times Month\ 1$	0.0002	−0.000049***	0.0029	0.0044	−0.0001	0.00004	0.0001
	(0.0004)	(0.000017)	(0.0063)	(0.0058)	(0.0003)	(0.00004)	(0.0001)
$Treatment \times Month\ 2$	0.0004	−0.000103***	−0.0080	−0.0045	0.0003	0.00004	−0.00003
	(0.0004)	(0.000020)	(0.0088)	(0.0072)	(0.0012)	(0.0001)	(0.00003)
$Treatment \times Month\ 3$	0.0001	−0.000102***	−0.0150**	−0.0071	−0.0007*	0.00004	−0.0000003
	(0.0004)	(0.000020)	(0.0067)	(0.0065)	(0.0004)	(0.00005)	(0.00004)
$Treatment \times Month\ 4$	0.0006	−0.000066**	0.0040	0.0031	−0.0010	−0.00001	0.0001
	(0.0005)	(0.000027)	(0.0104)	(0.0088)	(0.0006)	(0.00002)	(0.0001)
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	127,992,368	127,992,368	127,992,368	127,992,368	127,992,368	127,992,368	127,992,368
R^2	0.0018	0.000127	0.00002	0.00002	0.000003	0.000001	0.000001

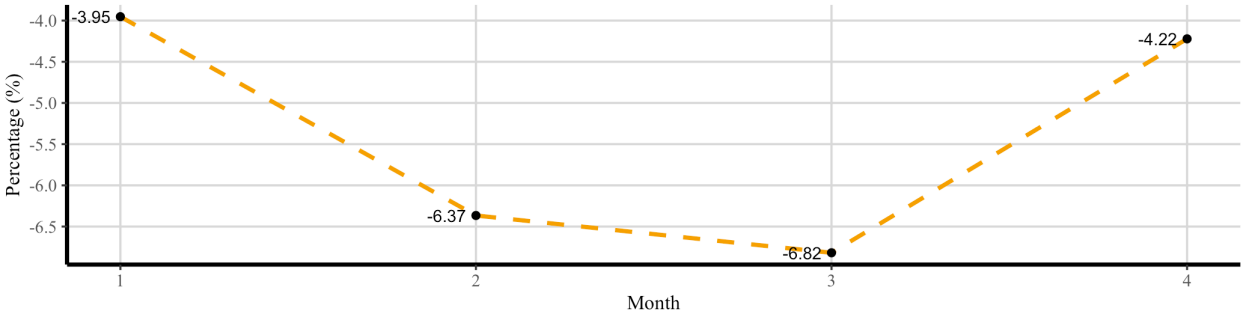
Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table D2: Time-varying Effects of Recommendation Diveristy on Inactive Users' Clicking Frequency

(a) Percentage Difference in Recommendation HHI Between Treatment and Control



(b) Percentage Difference in Clicking Frequency Between Treatment and Control



(c) The Recommendation HHI Elasticity of Clicking Frequency

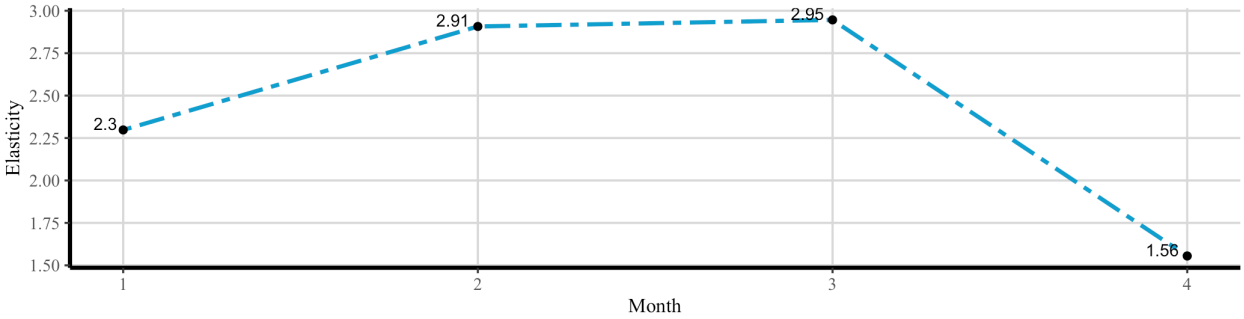


Table D3: The Manipulation Check and Treatment Effects of Recommendation Diversity for Active Users Over Time

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
$Treatment \times Month\ 1$	0.2479**	−0.0068***	0.0352***
	(0.1196)	(0.0007)	(0.0052)
$Treatment \times Month\ 2$	0.3330**	−0.0081***	0.0400***
	(0.1457)	(0.0007)	(0.0059)
$Treatment \times Month\ 3$	0.3275**	−0.0090***	0.0416***
	(0.1418)	(0.0007)	(0.0058)
$Treatment \times Month\ 4$	0.5410***	−0.0111***	0.0531***
	(0.1736)	(0.0010)	(0.0075)
Week Fixed Effects	Yes	Yes	Yes
Observations	1,604,026	1,604,026	1,604,026
R^2	0.0023	0.0110	0.0098

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
$Treatment \times Month\ 1$	0.3702	−0.0076	0.0313
	(0.2408)	(0.0063)	(0.0193)
$Treatment \times Month\ 2$	0.6087**	−0.0103	0.0357*
	(0.2875)	(0.0064)	(0.0203)
$Treatment \times Month\ 3$	0.5514**	−0.0137**	0.0474**
	(0.2706)	(0.0069)	(0.0212)
$Treatment \times Month\ 4$	0.6832**	−0.0216**	0.0741***
	(0.3143)	(0.0088)	(0.0260)
Week Fixed Effects	Yes	Yes	Yes
Observations	408,035	408,035	408,035
R^2	0.0019	0.0010	0.0015

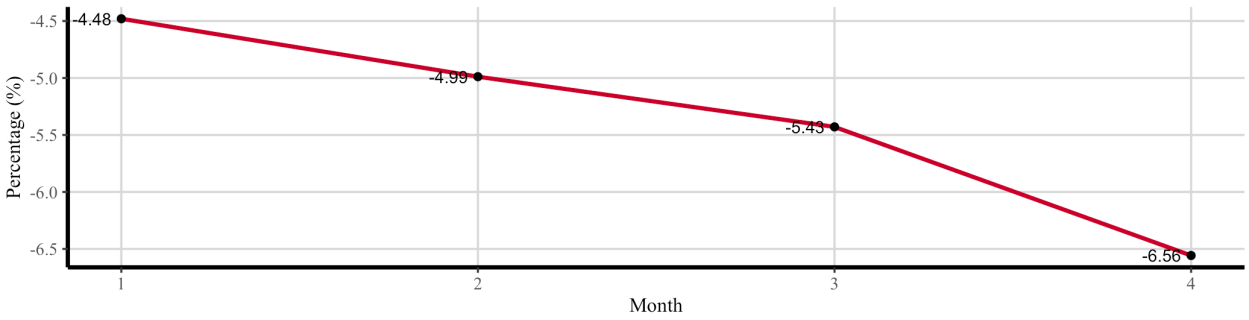
(c) Treatment Effect: Consumption and Engagement Levels

	consumption			*engagement*			
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
$Treatment \times Month\ 1$	−0.0003	−0.0007	0.1719	0.0551	−0.0201	−0.0007	−0.0007
	(0.0030)	(0.0008)	(0.3447)	(0.2578)	(0.0154)	(0.0005)	(0.0010)
$Treatment \times Month\ 2$	−0.0006	0.0002	0.7896*	0.5408	0.0137	−0.0003	0.00002
	(0.0030)	(0.0008)	(0.4278)	(0.3402)	(0.0229)	(0.0007)	(0.0011)
$Treatment \times Month\ 3$	0.0029	0.0003	0.3838	0.2913	0.0292	0.0001	−0.0006
	(0.0029)	(0.0007)	(0.2699)	(0.2158)	(0.0349)	(0.0008)	(0.0005)
$Treatment \times Month\ 4$	−0.0015	0.0004	0.5232	0.4675	0.0159	−0.0012**	−0.0001
	(0.0035)	(0.0008)	(0.3692)	(0.2939)	(0.0520)	(0.0006)	(0.0009)
Week Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,874,408	4,874,408	4,874,408	4,874,408	4,874,408	4,874,408	4,874,408
R^2	0.0075	0.0009	0.0003	0.0003	0.00004	0.00001	0.00003

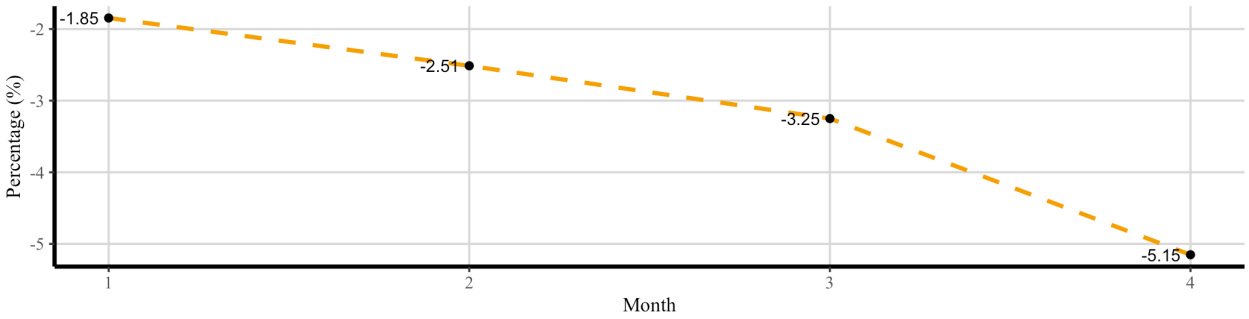
Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table D4: Time-varying Effects of Recommendation Diveristy on Active Users' Consumption Diversity

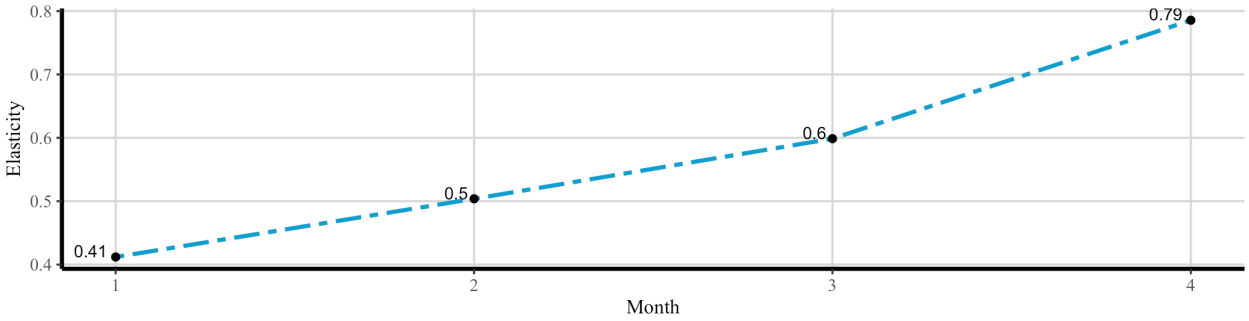
(a) Percentage Difference in Recommendation HHI Between Treatment and Control



(b) Percentage Difference in Consumption HHI Between Treatment and Control



(c) The Recommendation HHI Elasticity of Consumption HHI



Appendix E: NCM's Follow-up Experiment

NCM ran a follow-up experiment to test the benefit of increasing active users' recommendation diversity. The experiment lasted two weeks, from March 17 to March 30, 2022. Users of the control group received recommendations from the original recommender system, in which the window size was 5 for users who had clicked on videos in the past 30 days, and 15 otherwise. For the treatment group, NCM increased the window size to 15 for users who had clicked on videos in the past 30 days and decreased the window size to 1 for the other users. Consistent with our main analysis, we define active users as those who had watched videos four weeks before the experiment (i.e., from February 16 to March 15, 2022). Thus, active users of the treatment group would face higher recommendation diversity than those of the control group. We randomly sampled 1.5 million active users, and we show the experiment results in Table E1. The manipulation-check results confirm that treated active users were recommended 1.53% more video topics *(*$p = 0.0005$). Their recommendation HHI decreased by 3.16% ($p < 0.0001$) and their entropy increased by 1.15% *(*$p < 0.0001$).

Tables E1(b) and E1(c) show that the treatment significantly increased active users' consumption diversity as well as their consumption level without hurting engagement level. Specifically, treated active users clicked on 0.36 more topics ($p = 0.0002$, a 3.87% increase), decreased their consumption HHI by 2.42% ($p = 0.0002$), and increased their consumption entropy by 2.50% ($p < 0.0001$). Equivalently, a 1% decrease in the recommendation HHI led to a 0.77% decrease in the consumption HHI. Interestingly, active users in the treatment group also clicked on 0.59 more videos ($p = 0.0210$, a 6.62% increase) and spent 26 more seconds watching videos ($p = 0.0363$, a 5.87% increase) every week, which implies that a 1% decrease in the recommendation HHI resulted in 2.09% more clicks and 1.86% longer viewing time.

Table E1: The Manipulation Check and Treatment Effects of Recommendation Diversity for Active Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.1888***	−0.0054***	0.0234***
	(0.0544)	(0.0004)	(0.0024)
Week FE	Yes	Yes	Yes
Observations	2,051,928	2,051,928	2,051,928
R^2	0.0013	0.0003	0.0011

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.3647***	−0.0102***	0.0340***
	(0.0978)	(0.0027)	(0.0081)
Week FE	Yes	Yes	Yes
Observations	559,738	559,738	559,738
R^2	0.0004	0.0002	0.0003

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.0004	−0.0001	0.5872**	0.4286**	−0.0180	−0.0002	0.0013
	(0.0011)	(0.0006)	(0.2543)	(0.2047)	(0.0189)	(0.0009)	(0.0014)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,000,000	3,000,000	3,000,000	3,000,000	3,000,000	3,000,000	3,000,000
R^2	0.0014	0.0001	0.00004	0.00002	0.000002	0.000001	0.000001

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Appendix F: Comparing Two Measures of the Recommender Algorithm's Prediction Accuracy

We consider two possible indicators for measuring the algorithm's familiarity with a user: the number of clicks and the number of impressions a user had during the pre-experiment period. To find the more accurate indicator, we tested which one could better predict a user's future clicking-through rate. Specifically, among 2 million active users, we selected 203,263 high-valuation users in the control group who spent at least 10.42 minutes watching videos in the pre-experiment period and visited the Cloud Village in the first week of the experiment. Then we predicted their click-through rate (i.e., the number of clicks divided by the number of impressions) in the first experiment week by their standardized click (or impression) number during the pre-experiment period. Table F1 shows that the click number has a higher correlation with users' future click-through rate than the impression number (0.0726 vs. 0.0250). Consistently, the click number better predicts the future click-through rate in terms of R square (0.1390 vs. 0.0164). Thus, we choose the click number as the proxy for the prediction accuracy.

Table F1: Correlation Between Click-through Rate and Indicators of Algorithm's Prediction Accuracy

	Dependent variable:		
	click-through rate in the first week of the experiment		
	(1)	(2)	(3)
Standardized number of impressions in the pre-experiment period	0.0250***		−0.0266***
	(0.0004)		(0.0005)
Standardized number of clicks in the pre-experiment period		0.0726***	0.0881***
		(0.0004)	(0.0005)
Constant	0.1439***	0.1439***	0.1439***
	(0.0004)	(0.0004)	(0.0004)
Observations	203,263	203,263	203,263
R^2	0.0164	0.1390	0.1512

Notes: *p<0.1; **p<0.05; ***p<0.01.

Appendix G: Regression Tables for Each Segment of Active Users

Table G1: The Manipulation Check and Treatment Effects of Recommendation Diversity for High-Valuation, High-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.6180***	−0.0134***	0.0690***
	(0.1387)	(0.0004)	(0.0043)
Week FE	Yes	Yes	Yes
Observations	2,126,826	2,126,826	2,126,826
R^2	0.0031	0.0164	0.0122

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.3830***	−0.0080***	0.0342***
	(0.1313)	(0.0020)	(0.0076)
Week FE	Yes	Yes	Yes
Observations	1,219,555	1,219,555	1,219,555
R^2	0.0022	0.0014	0.0018

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0013	0.0008	0.7386	0.6262	0.0178	−0.0014	−0.0027*
	(0.0027)	(0.0013)	(0.5690)	(0.4594)	(0.0434)	(0.0015)	(0.0014)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,938,584	4,938,584	4,938,584	4,938,584	4,938,584	4,938,584	4,938,584
R^2	0.0102	0.0043	0.0020	0.0019	0.0003	0.00002	0.0001

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table G2: The Manipulation Check and Treatment Effects of Recommendation Diversity for High-Valuation, Low-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	−0.0216	−0.0102***	0.0410***
	(0.1999)	(0.0014)	(0.0102)
Week FE	Yes	Yes	Yes
Observations	213,936	213,936	213,936
R^2	0.0025	0.0097	0.0093

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	−0.2510	0.0174	−0.0407
	(0.2767)	(0.0110)	(0.0313)
Week FE	Yes	Yes	Yes
Observations	56,874	56,874	56,874
R^2	0.0033	0.0013	0.0020

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.0002	−0.0022***	−0.1894	−0.3205**	−0.0261***	−0.0015***	0.0018
	(0.0059)	(0.0008)	(0.1840)	(0.1570)	(0.0046)	(0.0003)	(0.0018)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	661,444	661,444	661,444	661,444	661,444	661,444	661,444
R^2	0.0072	0.0007	0.0001	0.0001	0.00004	0.00003	0.0001

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table G3: The Manipulation Check and Treatment Effects of Recommendation Diversity for Low-Valuation, High-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.4985**	−0.0092***	0.0485***
	(0.2425)	(0.0013)	(0.0106)
Week FE	Yes	Yes	Yes
Observations	266,290	266,290	266,290
R^2	0.0018	0.0089	0.0078

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.4706	−0.0146*	0.0463*
	(0.3030)	(0.0081)	(0.0256)
Week FE	Yes	Yes	Yes
Observations	92,485	92,485	92,485
R^2	0.0038	0.0008	0.0016

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	−0.0163***	−0.0002	0.4060	0.3240	−0.0189	−0.0010*	−0.0008
	(0.0056)	(0.0012)	(0.4118)	(0.3256)	(0.0189)	(0.0006)	(0.0011)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	764,036	764,036	764,036	764,036	764,036	764,036	764,036
R^2	0.0079	0.0012	0.0002	0.0002	0.0001	0.00002	0.0001

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table G4: The Manipulation Check and Treatment Effects of Recommendation Diversity for Low-Valuation, Low-Accuracy Users

(a) Manipulation Check: Recommendation Diversity

	num_recommended_topic	*HHI_recommended_topic*	*entropy_recommended_topic*
Treatment	0.2042***	−0.0069***	0.0322***
	(0.0306)	(0.0002)	(0.0017)
Week FE	Yes	Yes	Yes
Observations	6,596,126	6,596,126	6,596,126
R^2	0.0037	0.0111	0.0112

(b) Treatment Effect: Consumption Diversity

	num_clicked_topic	*HHI_clicked_topic*	*entropy_clicked_topic*
Treatment	0.0779	0.0022	0.0015
	(0.0714)	(0.0028)	(0.0077)
Week FE	Yes	Yes	Yes
Observations	975,996	975,996	975,996
R^2	0.0027	0.0013	0.0019

(c) Treatment Effect: Consumption and Engagement Levels

	consumption				*engagement*		
	visit	*freq_click*	*num_click*	*view_min*	*num_like*	*num_comment*	*num_share*
Treatment	0.0003	−0.0002**	0.0217	0.0151	−0.0022	0.0002	−0.00002
	(0.0010)	(0.0001)	(0.0323)	(0.0251)	(0.0018)	(0.0002)	(0.0001)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	21,635,936	21,635,936	21,635,936	21,635,936	21,635,936	21,635,936	21,635,936
R^2	0.0071	0.0003	0.00004	0.00003	0.00001	0.00001	0.00001

Notes: Standard errors clustered at the individual level are in parentheses. *p<0.1; **p<0.05; ***p<0.01.