Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Altering Words to Evade Perceived Moderation: Decoding Algospeak in Chinese Social Media Video Captions

Andy Zhao
Cornell University
Ithaca, New York, USA
dz352@cornell.edu

Abstract

Video content creators on social media often modify their caption text to bypass algorithmic moderation as they perceive it – a practice known as algospeak. This behavior of adjusting expressions online could become especially complicated in Chinese online space where both moderation and censorship overlap, and where certain types of expression are explicitly restricted. In this study, we conduct a large-scale analysis of algospeak by examining over 200,000 movie-recap videos from 96 YouTube channels, most of whom are also prominent on Douyin (the domestic implementation of TikTok in China). We extract the video's hard-coded captions and employed AI to detect altered terms. We then examine the prevalence and characteristics of these alterations, and leverage two commercial text-sensitivity detectors to infer the potential motivations and policies driving these algospeak behaviors. We also explore whether factors such as content reach and monetization strategies correlate with creators' propensity to modify captions. We find that the presence of embedded ads in videos as a form of brand deals was negatively associated with overall alteration rates: content creators increased alterations for a few explicitly warned topics while reducing them for other topics. Overall, we argue that algospeak in Chinese videos represents a collectively adopted strategy that both evades perceived moderation and reinforces restrictive norms within online communities, and is partly attributable to a trust gap between users and the platform on the matter of moderation.

Introduction

Many social media platforms rely on complex algorithms to both recommend and moderate contents. However, these systems are usually not transparent enough for ordinary users without technical backgrounds to understand (Felaco 2025). Many users believe that the algorithms could restrict certain topics in unpredictable ways (Lorenz 2022). To avoid undesired moderation, some users will tweak their expressions or use euphemisms—a phenomenon referred to as algospeak (Klug, Steen, and Yurechko 2023; Felaco, Pelliccia et al. 2024). In social media videos, a common algospeak strategy is to alter the caption text, rather than the spoken audio, by replacing some characters with semantically or phonetically comprehensible variants, such as changing "die" to "unalive" or "sex" to "seggs" (Steen, Yurechko, and Klug 2023). This strategy is perceived as a linguistic innovation to adapt the algorithmic moderation on audiovisual platforms like TikTok (Calhoun and Fawcett 2023).

In Chinese digital space, using algospeak to evade the moderation is further complicated by censorship practice in China that employs strict keyword-based restriction, particularly on political expressions (Fu, Chan, and Chau 2013). Chinese netizens long ago developed a rich corpus of euphemisms to circumvent censorship decades ago (Lee 2016). Nowadays, these historical practices of circumventing keyword-based censorship encountered with algorithmic moderation on social media, creating a hybrid landscape in which platform governance and government regulations operate together. As a result, using coded words on the Internet has shifted from a resistance tactic of the few to a common and routine behavior adopted by more people that extends well beyond politically sensitive topics (Bao 2025). This evolution raises new questions about the Chinese algospeak in terms of its prevalence, characteristics, and associated or driving factors, particularly given the distinctive context in China. Answering these questions could reveal how platform norms and sociopolitical contexts contribute to diverse and nuanced digital behaviors, even when similar behaviors emerge across global online spaces.

While earlier studies have relied on qualitative methods such as interviews to explore algospeak practices, quantitative analysis of algospeak has proven challenging. The difficulty largely comes from the complexity of identifying the nuanced linguistic modifications across different media formats, as well as the lack of large-scale video datasets. However, with the rise of cross-posting behaviors on social media platforms such as Douyin and YouTube (Meng and Nansen 2022), there is now an opportunity for large-scale, systematic analysis of algospeak, both in terms of its prevalence and its nuances, within the more restrictive regulatory and linguistic environment of the Chinese online space.

In this study, we collected over 200,000 videos about "movie-recap" from YouTube channels, most of whom also posted on Douyin, and extracted hard-coded captions using computer vision techniques.[1] Figure 2 in the Appendix

[1]While we could not provide a statistical description about cross-posting behaviors given the data access restriction on

presents an example of word alteration in a movie-recap video caption, where the word 精灵 (“elf”) was altered into 精L. We then leveraged AI to detect possible word alteration in the captions and built a dictionary of algospeak terms to facilitate our analysis. To understand the algospeak in the Chinese digital space, we aim to address three key research questions:

- RQ1: How prevalent is word alteration in the captions of Chinese social media videos, and what characteristics define this phenomenon?
- RQ2: What exogenous concerns drive content creators' decisions to modify their content?
- RQ3: Are content creators' decisions to adjust their own works associated with factors such as creator reach and monetization strategies?

In this paper, we find that (RQ1) word alteration is a common practice for around 80% of video content creators. The terms associated with violence or crime are usually heavily altered, and superlative and absolutist expressions, like “first” or “most”, are often intentionally avoided. We then leverage third-party commercial tools, which were designed to flag sensitive terms with explanations, to infer that (RQ2) the altered expressions are discouraged because they usually contain uncivil and violent languages, or violate the advertising law (a law in China that explicitly prohibits absolute expressions).We also find that (RQ3) monetization through embedded ads is surprisingly negatively associated with video alteration rates, instead of incentivizing creators to adopt more cautious or risk-averse content strategies. Specifically, content creators increased the alteration rates in the explicitly-warned topics while altered less in other categories in the videos with embedded ads, as a form of brand deals, compared to other videos.

Overall, we argue that what is important in shaping algospeak behavior is what users believe about moderation, rather than what is actually moderated, and a trust gap between users and the platform make the content creators to continue altering their expressions regardless of the platform assurances. Our findings suggest that the platform design influences not only what content remain online, but also how users learn to communicate, with long-term implication for trust, norms, and creation. We interpret the algospeak as a collective strategy of users, extending beyond the specific context of Chinese social media, to evade moderation and reclaim control over works but also reinforce a restrictive norm in the online communities.

Related Work

This work builds on the previous studies from multiple domains. We first introduce the studies about algorithmic moderation on social media which sets the background of user perception and behaviors. Next, we introduce the works about existing algospeak behaviors, which are most relevant to our topic. Last, we introduce the researches about the censorship and moderation on Chinese audiovisual contents, which provides a background about Chinese online space.

Algorithmic Moderation

Social media platforms rely heavily on machine-learning and neural-network algorithms for governance and feed curation (Kim 2017). These automated moderation systems can make seemingly arbitrary decisions that largely shape user experiences and welfare (Gomez et al. 2024). Because these powerful algorithms operate as opaque “black boxes,” their lack of transparency only deepens concerns about an already secretive moderation regime (Roberts 2019; Suzor et al. 2019). In the eyes of many users, algorithmic moderation, the fundamental architecture on these platforms, is opaque and difficult to comprehend or audit (Gorwa, Binns, and Katzenbach 2020; Cotter 2023). Many studies suggested that social media users often find themselves in a vulnerable position, feeling confused and frustrated by the absence of clear explanations, accountability mechanisms, and appealing channels, and experiencing unfair or inconsistent treatments (Vaccaro, Sandvig, and Karahalios 2020; Zeng and Kaye 2022; Ma et al. 2023; Abokhodair et al. 2024).

The lack of transparency and the agency in governance erodes users' trust about social media and its algorithmic moderation (Brunk, Mattern, and Riehle 2019; Molina and Sundar 2022). A sense of control over their content is important for users to main their autonomy and form their trust on the platform, but the arbitrary moderation decisions could undermine that confidence and leave users feeling powerless (Naher, An, and Kim 2019; Vaccaro, Sandvig, and Karahalios 2020; Hödl and Myrach 2023; Ma et al. 2023).In the face of huge uncertainty, users often struggle to get a sense of the system and consequently develop their informal folk theories about how the algorithm works, though these beliefs are usually not validated (Eslami et al. 2016; DeVito et al. 2018; Savolainen 2022). Many folk theories attempt to explain content removals or income decreases, especially among users who feel targeted or belong to marginalized groups (Mayworm et al. 2024).Based on these beliefs, users will start to either adapt their language and behaviors to appease the algorithm to gain more viewership, or they question and resist the algorithmic curation (Vaccaro, Sandvig, and Karahalios 2020; Lin 2025).

Algospeak

Algospeak is a prominent way that users take to contest the algorithmic moderation and widely adopted by users on different platforms (Lorenz 2022). Chancellor et al. (2016) found that Instagram users would use different lexical variations to avoid the algorithmic restriction on the discussion about eating disorders, and the online communities adapted such strategies grew even stronger. TikTok is another primary field associated with algospeak since content creators greatly relies on the algorithmic curation on this audiovisual platform (Guinaudeau, Munger, and Votta 2022). Steen, Yurechko, and Klug (2023) interviewed TikTok content creators and found that the primary motivation of algospeak was to evade algorithmic moderation, which they perceived as inconsistent and biased, and reach their audience.

Douyin, we have an example to illustrate this common phenomenon in Appendix.

Dawson (2024) examined the exact strategies that users take in algospeak and categorized different euphemism strategies in algospeak into three categories: tactical misspellings in the words, using the emojis as substitutions in the text, and a performative word evasion which completely use another term to refer a more sensitive topic (like "unalived" for "death" in a Orwellian Newspeak way).

The algospeak behavior does not evenly impact all topics. Content creators' moderation evasion behaviors were usually related to the topics like sex and sexuality, or sometimes offensive dog whistle words (Calhoun and Fawcett 2023; Dawson 2024), and could backfire on the quality of contents (Klug, Steen, and Yurechko 2023).

Some studies attempted to investigate how the algorithm awareness could impact algospeak (Klug, Steen, and Yurechko 2023; Felaco, Pelliccia et al. 2024). Though not conclusive, Felaco, Pelliccia et al. (2024) still indicate that the algospeak is closely tied to users' algorithm literacy and experience with algorithmic moderation.

Censorship and Moderation on Chinese Audiovisual Contents

Chinese digital space is infamous for its strict information control and various methods of censorship, and netizens adapt to this environment with various of self-censorship or censorship circumvention behaviors (Lee 2016; Chang et al. 2022; Chen et al. 2023). Self-censorship and proactive manipulation in audiovisual contents from users, especially the subtitles or captions like our subjects, are very prevalent in Chinese cyberspace (Wang 2020). Yan (2025) found that grass-root movie translators would use their taken-for-granted beliefs about sensitivities, even without direct governmental control or explicit policies, to adjust their works, often with varying levels of self-censorship across different teams. She suggested that productive censorship, the artful evasion of content restriction (Baer 2011), could benefit the marketability of audiovisual content.

Moreover, it is noteworthy that content moderation is still an organic part of Chinese social media platforms, even though it could be intertwined with the censorship (Li and Zhou 2024; Zhao and Hu 2025). Douyin, as one of the most popular audiovisual platforms in China, even attempt to leverage the power of voluntary users to improve its platform governance in a more participatory way (Ye, Huang, and Krijnen 2025).Content moderation on Chinese platforms is a practice as vigorous as their Western counterparts (Zhao and Hu 2025). Although it is true that social media platforms have to enforce the governmental policies on the online contents and take more proactive actions to restrict various expressions (China Media Project 2021).

Methods and Data

In this section, we first explain how we collected the video data from social media and extract their hard-coded captions. Then, we introduce the process to build the dictionary of altered terms, which is essential to identify all word alteration cases. Next, we talk about how we categorize the words in our dictionary of altered terms, and then the movie caption text. Last, we explain the process to detect the embedded ads in the movie captions.

Data Collection and Captions Extraction

To begin with, we want to collect the videos from social media, where algospeak usually happens. We searched all videos that contains the Chinese word "movie-recap" (电影解说) on YouTube and identified 96 channels that had posted such videos and had at least 1000 videos in total. We focused on "movie-recap" videos for several reasons. First, they are highly text-rich, because creators need to narrate and usually caption the videos. Their untouched audio data and altered caption text serve a good dataset to identify algospeak behaviors. Second, movie recaps videos usually have contents about violence and sexuality that are more likely to trigger automated moderation than genuine topics, and increase the chance for algospeak use. While caption contents are centered on movies, their thematic focus nevertheless covers a wide range of topics that are relevant to social media moderation. Such topic focuses indicate that "movie-recap" videos would cover a wide range of sensitive topics, though heavily movie-themed. At the same time, these "movie-recap" videos are popular on Chinese internet because the recapped movies are usually not licensed in China, so they serve as a popular channel for social media users to know about those unlicensed movies. This structural constraint may further raise creators' awareness of moderation risks and may amplify algospeak strategies. By contrast, other popular genres, such as comedy clips, pet, and education videos, tend to feature less extensive caption text and are less likely to use sensitive terms. As a result, algospeak in these genres may be less frequent or more narrowly distributed compared to "movie-recap" videos. Therefore, we believe that movie-recap videos are especially likely to exhibit both stronger and more diverse forms of algospeak among various content genres, making them a great fit for our empirical analysis of the algospeak phenomenon.

Also, we chose YouTube instead of other short-video platforms like TikTok or Chinese platforms like Bilibili because YouTube is more friendly to video data collection. In addition, many Chinese content creators would upload their works in multiple platforms, even outside the Great Firewall. In fact, we identified 73 of 96 YouTube channels also operated on Douyin with the same name and the same contents but a much popular base. Only 27.6% of these channels on YouTube were created before 2024. It is also noteworthy that these accounts have simplified Chinese captions, but many of their titles or descriptions on YouTube are in traditional Chinese. This existence of multi-Chinese contents indicates that the creators of these movie-recap are probably mainland Chinese but the audience are likely overseas Chinese-speaking users, and these content creators were re-posting their works from other platforms.

Next, we retrieved all videos from 96 channels, and identified all videos posted before 2025 that contain the simplified or traditional Chinese term "recap" (解说) in the title or description. We eventually obtained 200,913 videos and downloaded them with Python package "yt_dlp".

Then, we leveraged an open-sourced computer vision tool called video-subtitle-extractor [2] to automatically extract the hard-coded captions from these movie-recap videos. This tool is primarily built on the library OpenCV for caption detection. We optimized this tool for speed and also customized it to adjust the videos of different frame sizes.

Build the Dictionary of Altered Terms (DAT)

There are multiple algospeak practices in Chinese online space, similar to TikTok algospeak (Dawson 2024). First, content creators could replace the Chinese character with Latin alphabets in Pinyin (a Romanization pronunciation system), which could be treated as a tactical misspelling. In a similar English example, the term "police" could be changed to "p0l1ce" in the text to avoid algorithmic detection. Second, content creators could intentionally split or even skip the perceived sensitive terms in the caption in a way that does not affect the comprehension. Third, content creators could replace some common terms with emojis or other expressions that may be comprehensible to most audience (referring to the police as the Uncle Hat). Given the difficulty to detect the other two cases in a clean way, our study only focuses the first condition. The Chinese algospeak cases are operationally defined as scenarios in which Chinese characters are replaced by Latin (Pinyin) alphabets in a way that is not linguistically necessary. Thus, our findings represent a lower bound of Chinese algospeak phenomenon.

To trace and identify the potential altered terms, we built a dictionary of altered terms (DAT) to record the relations between altered terms and their original terms. In our dictionary, we only considered altered Chinese terms that meet at least one of the following criterion: at least one Chinese character remained after alteration (eg. 麻药 and 麻y), or at least three alphabets in the altered form (公务员 and gwy), or full Pinyin spelling instead of just initials (eg. 嫌疑 and xianyi). These three criterion provides a relative high confidence to anchor the original terms without manually checking the videos all the time. By cataloging which words get altered, we essentially aimed to reverse-engineer the blocklist the content creators have in mind.

Our dictionary was built with the assistance of GPT-4o, because manually going through all captions is too demanding and LLM is proven to perform in understanding the algospeak (Fillies and Paschke 2024). We employed a GPT-4o model with few-shot prompting (shown in the Appendix), using manually curated examples in real video captions to illustrate the inference of algospeak terms in Chinese, and then fine-tuned this model with 20 more different algospeak examples and expected answers. Given a raw caption text as input, our model was instructed to identify potential algospeak terms with Pinyin alphabets and infer their most likely original terms. For each caption, our fine-turned model returned structured inference results consisting of the identified algospeak terms present in the caption and corresponding inferred original terms. We then carefully examined all responses generated by the GPT-4o model and kept only those plausible altered terms for which we could confidently verify the inference results, either directly when the interpretation was unambiguous or by validating them against the original context when multiple interpretations were possible. This manual step involved retrieving the highly possible combination of Chinese characters and words based on the Pinyin hints and cross-checking some suggested altered terms against the original full caption text or video content to assess their plausibility. While we could not fully exclude cases where a single algospeak term plausibly corresponded to multiple original terms (e.g., s人 could plausibly refer to both 杀人(kill a person) and 死人(dead person)), our manual examination of every entry guarantees high precision. Because our goal is to characterize norms rather than exhaustively enumerate all algospeak examples, we prioritize precision over recall (which we could not provide a robust estimation). Eventually, our approach eventually resulted in a dictionary of 2,067 algospeak terms.

Categorizing the Altered Terms and Movie Captions

To reveal the topic focus of Chinese algospeak and facilitate our analysis, we planned to identify the topics among altered terms. We first embedded all unique original terms in our dictionary into vectors in a 1536-dimensional space with OpenAI's text-embedding-3-small model.

Next, rather than relying on fully unsupervised clustering, we chose a semi-supervised, centroid-based assignment procedure. In preliminary analyses, we experimented extensively with K-means clustering using a range of cluster numbers. Although several topical groupings consistently emerged across runs, the resulting clusters were unstable in size and sensitive to hyperparameter choices, making it difficult to obtain a robust and interpretable result.

Therefore, we leveraged the recurring topic from our K-means runs as the centroids in each topic group. For each theme, we selected a representative keyword to serve as a semantic proxy for the cluster centroid. All terms in DAT were then assigned to the topic whose centroid keyword was closest in embedding space, using cosine similarity. This approach kept the recurring topical structure observed in unsupervised clustering and ensured stable, interpretable, and balanced topic categories.

Our topic keyword selection was consistent with what commercial detector suggested, which we will show later. The topics and the number of unique terms in each topic category are shown in Table 1.

Additionally, there are two special groups in our result. The first group is "Most". We found that many terms like "most" or "first" got altered in the captions (which we will explain in our results) but the meanings of these terms vary too much. So we excluded these terms in the clustering step and assigned them into a single group called "Most". The second group is "Others". Because some miscellaneous terms could not be clustered in any group appropriately. We calculated the average distance between all words and their closest centroids, along with the standard deviation of these distances. Terms whose shortest distance to any centroid exceeded the sum of the average distance plus one standard deviation were assigned to this "Others" group.

[2] https://github.com/YaoFANGUK/video-subtitle-extractor/tree/main

Topic	DAT words	Altered	Not Alt.	Alt. Rate (%)
Kill	206	20802	356856	5.51
Death	226	18097	542289	3.23
Most	99	10082	554101	1.79
Police	125	8391	255548	3.18
Weapon	92	6022	167343	3.47
Poison/Drug	112	4693	105741	4.25
Blood	140	3495	140017	2.44
Ghost	131	3353	104972	3.10
Crime	143	3241	211921	1.51
Others	174	2783	243936	1.13
Money	130	2677	240813	1.10
Luxury	80	1898	65250	2.83
State	38	1599	89141	1.76
Sex	59	1586	58269	2.65
Medical	46	1046	87713	1.18

Table 1: Alteration by Topic: this table presents, for each topic group, the number of altered versus non-altered cases, and the number of DAT words in each topic.

Detect the Embedded Ads

One research question involves the monetization of content creators, and one monetization strategy is to insert an advertisement clip in the middle of a video. These advertisement clips are not associated with the platforms like YouTube, which may inject the ads anywhere in the videos. Rather, these ads are brand deals made by the creators and embedded within the movie-recap contents, recited by the creators themselves and displayed in the captions. For instance, while describing a car chase scene, the video presenter suddenly shifts to promoting a pre-owned car marketplace. Figure 3 in the Appendix displays an example of embedded ad.

However, there are all kinds of advertisements and they do not follow a fixed pattern of exclusive keywords. To identify these embedded ads, we need to build our own comprehensive list of keywords or expressions for our movie-recap videos. We started with the keywords like “download”, “link”, “comment section”, “left bottom corner”, and “recommend”, and some common brand names. We detected some ads in the captions and constructed exclusive expressions to retrieve these ads. Then we randomly examined 50 more videos from each channel that had embedded ads, which returns more distinctive ads in other videos and their corresponding exclusive expressions. With new ads-related expressions and variations of brand names in the ads, we repeated the above steps for multiple rounds, till we could not find more ads in a new round of two hundreds of randomly-retrieved videos in all channels. In the end, we created a list of 23 different expressions of ads, and found 26 channels ever inserted ads in their videos, though only 498 (0.3%) of all videos have embedded ads.

Results

We first present descriptive results about the word alteration in Chinese movie-recap videos, which provides a comprehensive understanding in terms of topics, frequencies and trends about this phenomenon. Next, we present the response results from commercial sensitivity detectors and combine their proposed reasons with our alteration data. We then infer the possible reasons that could convince the content creators to alter words in the captions. Last, by crossing the collected YouTube channels with Douyin account data and retrieving channel-level followers and likes information, we examined whether the word alteration in the videos are associated with reach and monetization of content creators. We specifically explored the impact of the first embedded ads and the uniqueness of the videos with embedded ads in algospeak practices.

Prevalence and Characteristics of Algospeak in Chinese Movie-recap Videos

We investigate the prevalence of algospeak in Chinese movie-recap videos based on over two thousand words in our dictionary. 79.83% of all channels have algospeak practice, and 15.14% videos have altered terms in their videos. We computed the word alteration rates on both video-level and channel level. This alteration rate is defined as, out of all words in our dictionary, how many of them were used in an altered format, compared to all occurrences of these words in any format in the video captions. We find around 9.3% of all videos and 30.3% of all channels had over 10% alteration rates (Figure 5 in Appendix), suggesting that word alteration behaviors were concentrated in a small group of all content creators of movie-recap videos.

Because the word usage of different topics are not uniform in the movie-related videos, we present the word alteration by topics and the word occurrences in the original forms in Table 1. The topic-level alteration rate is defined as all words of a certain topic appeared in an altered form divided by the all words of this topic appeared in any form. Table 1 indicates that the most altered group is “Kill”, which has the highest alteration rate as 5.51% and have the most alteration cases (20,802) in our dataset. The next two most altered groups are “Poison/Drug” (4.25%) and “Weapon” (3.47%). The words about “Death” have the second largest alteration cases (18,097), but their alteration rate only ranked fourth (3.23%). The group “Most”, which contains superlative and absolutist expressions like “first” or “most”, does not necessarily have a high alteration rate (1.79%) even though it has the third largest alteration cases (10,082).

To further explore the pattern of most altered words in the movie-recap video captions, we dive into the top 100 most altered words. Figure 1 presents these words in terms of their word-level alteration rates (Y-axis) and total occurrence numbers in log scale (X-axis). Each dot represents a word and the dot color represents their topic categories, and some dots are associated with their original term for better interpretation. Some orange or brown dots, which represent the words about drugs and violent scenes, locate on the top left corner of Figure 1, suggesting that they were heavily altered even though their total occurrences are relatively low.. On the other hand, the ordinal terms like “first” and “last” (green dots) stay in the right bottom corner of Figure 1, indicating that they have low alteration rates even though the total occurrence numbers are high.

Additionally, we find that over 75% channels altered terms in most topic groups, only with three groups as excep-

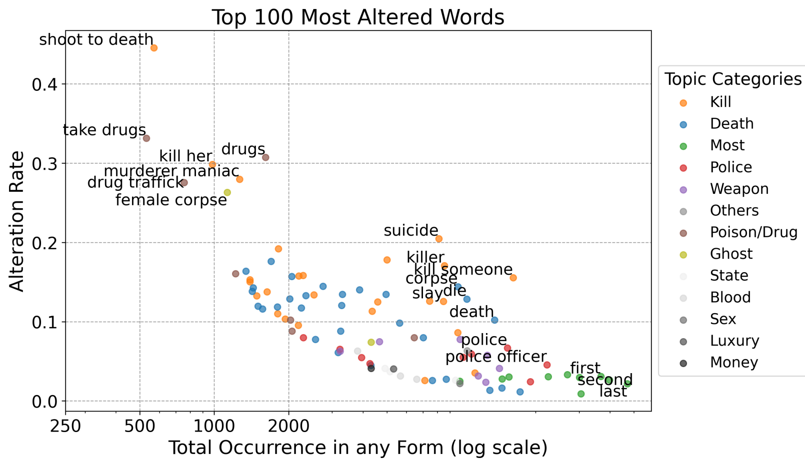


Figure 1: Scatterplot of top 100 alter words. The x-axis represents the total occurrence of each word in any form (original or altered) and the y-axis represents the alteration ratio of each word (altered occurrence divided by the total occurrence). The dot color represents the topic category and some prominent dots are also labeled with their word text.

tions of low consensus for alteration: Ghost, Medical, and Luxury topics. These collective and nearly unanimous alterations suggest that video creators may hold a shared belief that certain topics are heavily targeted by platform moderation and therefore require extra caution.

Inferring the Concerns Behind Chinese Algospeak

TikTok users suggested that their algospeak behaviors resulted from their anticipation about algorithmic moderation bias (Steen, Yurechko, and Klug 2023), and we extend this insight by leveraging third-party resources to infer content creators' potential motives. We take advantages of two popular commercial text sensitivity detectors (Cizhua and Check51), which were used by content creators to identify the phrases that might trigger moderation before posting their works. We reversed our caption text with our dictionary into their original text, randomly sampled 5% text from all data, and prompted these two tools to flag terms in the captions. Both tools responded with not only the flagged terms but their assumed reasons to discourage the use of these terms. We do not claim that these flagged terms are necessarily altered in real captions; rather, we use detector responses as a proxy for creators' beliefs about moderation sensitivity. Also, we acknowledge that commercial detectors may overstate the severity of certain terms in order to demonstrate their usefulness and reduce the risk of missed detections, or may fail to identify relevant cases due to their limits in detection algorithms. Most suggestions across various warning levels, 947 flags from Cizhua and 3116 flags from Check51, were not really altered by content creators. In comparison, among the 171 real algospeak cases in these sampled captions, only 61 altered words (35.7%) were detected by at least one of the two detectors, while the remaining 64.3% (110 altered words) were not detected. So, we only carefully use the responses of commercial detectors as a third-party reference instead of a golden rule, and the reasons of these two tools are often different even for the same term.

Table 2 presents the top five reasons content creators altered their video captions from the perspectives of both Cizhua and Check51. The most common reason from Cizhua is about the uncivilized and violent language (21.21%) and the most common reason from Check 51 is about the advertising law (in China) or precedent cases (32.74%). We also associated the real alteration in captions with Cizhua and Check51's reasons and presented a more comprehensive result in the Table 6 in Appendix. While Table 2 and Table 6 do not represent the true thoughts from the content creators, the reasons behind altered terms from both tools are largely consistent and converge on a common "forbidden zone" in video content: no absolute expression, no uncivil or overtly negative expressions. This consensus between commercial sensitivity detectors reflects the shared knowledge within the content creation community about what could be allowed and what are discouraged.

As a matter of fact, video platforms like Douyin are aware of the existence of algospeak and not supportive of such behaviors. Douyin have explicitly gave public instructions to their e-commerce users about the allowed and forbidden language usage, and even provided a list of discouraged algospeak examples (Douyin Blackboard 2022). Many instructions from Douyin were actually consistent with the reasons mentioned by Cizhua and Check51, especially avoiding the absolutist expressions and being careful about medical-related contents. In addition, Douyin's vice president Liang Li informally explained that the restriction is only limited to e-commerce context and not imposed on the ordinary users, and he also said:

> *Terms like 米 (mi)[3] are just rumors being spread around. There are many so-called "Douyin operation guides" on the market that list numerous supposed sensitive words, which people believe without question. We've debunked these rumors before, but many people remain unaware.*

[3] 米 (mi) is commonly used to refer "money".

	Cizhua			Check51		
	Reason	Percent	Example	Reason	Percent	Example
1	Uncivilized and violent language	21.21%	Kidnap	This term may violate advertising law or has appeared in known penalty cases, please make your own judgment.	32.74%	Call the Police
2	Involving politics, religion, race, gender, terrorism, etc., affecting social stability	7.57%	White House	Suspected of Containing Insulting, Obscene, Pornographic, Gambling, Superstitious, Terroristic, Violent, and Similar Content	10.44%	Revenge
3	Suspected of Containing Insulting, Obscene, Pornographic, Gambling, Superstitious, Terroristic, Violent, and Similar Content	6.95%	Underwear	Words suspected of being absolute, extreme, and impossible to verify	10.34%	Only
4	Involves violence, endangers personal safety	6.54%	Domestic Violence	Feudal (reactionary), superstitious, pseudo-scientific words	9.25%	God
5	Extreme Words/Absolutist Language; Evidence Must Be Provided to Prove Conformity with Facts, Otherwise It Is Considered an Absolute Statement	5.74%	Famous	Exaggeration Terms, when Used in Advertisements, One Cannot Exaggerate a Product's Leading Position or Superiority to Avoid Misleading Consumers	5.49%	First

Table 2: Top Five Reasons Proposed by Cizhua and Check51 for the Terms Altered in Video Captions. The "Percent" column shows how frequently each tool identified this particular reason out of all its responses. The 'Example' column provides a common example that falls under this category.

Assessing the Relationship Between Algospeak, Creator Reach, and Monetization

Most creators of movie-recap videos are driven by business interests and are therefore expected to be risk-averse toward factors that may impede video reach or audience growth. As a result, algospeak behaviors, which could be understood as strategic responses to perceived moderation risks, may be associated with creators' monetization status and audience reach. To examine this potential association between creators' endogenous attributes and their algospeak behavior, we focus on factors such as creator reach and monetization. We hypothesize that creators who have profited from their videos have more to lose from unfavorable moderation decisions and therefore have stronger incentives to adopt word-alteration strategies.

Meanwhile, we noticed that the YouTube channels which also existed on Douyin usually have much more followers and videos, and also longer operation time on the latter platform than those of their YouTube accounts. So, we believe that YouTube is not the primary field of many content creators, and we choose to use the channel information (total likes and followers) of the matched content creators on Douyin as a more accurate description about these creators (thus with a smaller sample than the previous analysis). Regarding to the monetization, creators could earn compensation through the views of their videos from the Douyin or YouTube. Since monetization is already correlated with the content reach to some degree, the total likes and followers of these creators can function as a proxy of monetization. However, we have included another variable of monetization in our analysis: the embedded ads, because it is not uncommon for content creators to insert an advertisement clip into their video and prompt a product. We have explained how we identified the videos with embedded ads in the section .

With our YouTube caption and Douyin data, we prepared channel-level data – total likes and follower numbers – and video-level variables: a binary variable of "embedded ads", the length of video captions, and the alteration rate. Formally, we estimated a hierarchical regression with the alteration rate as the dependent variable. Video-level alteration rate was regressed on both video-level predictors (embedded ads, caption length) and channel-level predictors (number of followers and likes) were included as fixed effects, with a random intercept specified for each channel. This model specification allows baseline alteration rates to vary across channels while assuming common slopes for the covariates. All continuous predictors were log-transformed to reduce skewness and improve interpretability.

The model was estimated using maximum likelihood, as implemented in the MixedLM function in statsmodels in Python. By replacing channel fixed-effect dummies with random intercepts, the model avoids perfect collinearity that would otherwise arise when time-invariant channel-level variables (e.g., followers and likes) are absorbed. Moreover, partial pooling improves estimation efficiency by reducing the influence of random noise from channels with few observations by shrinking their estimates toward the overall mean. We report coefficient estimates in percentage-point terms for ease of interpretation, holding other variables constant.

Table 3 presents the regression results, where Model 1 uses the alteration rate of all terms as the dependent variable. Surprisingly, our result suggests that whether a video contains an embedded ads has a significantly negative, not positive, correlation with the video alteration rates, which is the opposite of our hypothesis. On average, having an embedded ad in the video will decrease the alteration rate by 1.26%, holding other variables constant. The length of captions (logged form) has a significantly positive correlation with the alteration rate, suggesting that, for example, doubling the length of a caption of 1,000 characters will increase the alteration rate slightly by around 0.25%. The number of likes and followers (logged form) both have no significant correlation with the video alteration rates, since their p-values are bigger than 0.05. We then narrow our analysis down by using the alteration rates of 189 terms that were altered at least 100 times in all videos. Model 2 in Table 3 represents our analysis with the alteration rate of only frequent terms as the dependent variable. Compared to Model 1, the general trend of our result does not change but the coefficient of "ads" increased greatly. Controlling other variables, the alteration rates of those 189 frequent terms in the videos will increase by 3.11% with an embedded ad compare to a video without ads. The effect of the caption length also raises by a little – for a 1,000-character caption, doubling its length increases the alteration rate by about 0.68%.

Statistic	Model 1	Model 2
Dependent Variable	`Alter Rate`	`Alter Rate (frequent terms)`
No. Observations	142,036	142,036
No. Groups	70	70
Scale	64.0003	246.8880
Min. group size	140	140
Log-Likelihood	-497130.8950	-593005.9704
Max. group size	4115	4115
Mean group size	2029.1	2029.1

	Model 1			Model 2		
Variable	Coef (SE)	z	P> |z|	Coef (SE)	z	P> |z|
Intercept	-1.260 (2.610)	-0.483	0.629	-4.930 (5.186)	-0.951	0.342
Ads	-1.262 (0.374)	-3.371	0.001	-3.109 (0.735)	-4.229	0.000
Length (log)	0.353 (0.019)	18.505	0.000	0.978 (0.037)	26.097	0.000
Followers (log)	-0.183 (1.068)	-0.171	0.864	-0.287 (2.123)	-0.135	0.892
Likes (log)	0.427 (0.833)	0.512	0.608	0.847 (1.655)	0.512	0.609
Group Var	28.805			113.746		

Table 3: Regression Results of Two Mixed Linear Models on the Alter Rates. Model 1 has alteration rate of all terms as the dependent variable, and the Model 2 has alteration rate of only the frequently altered terms as the dependent variable. Both models suggest that only the existence of embedded ads in the videos and the caption length have significant implication on the video-level alteration rates.

We have also attempted to explore whether the introduction of the very first embedded ads had changed the channel-level alteration patterns, but our analysis received null results (as shown in the Appendix). Therefore, to gain more insights about the unique relation of embedded ads and word alteration, we investigate the alteration rates and word occurrence across various topics and video types. We compare three types of videos in our dataset: videos with embedded ads (the group "Ad"), non-ad videos posted on the same day as ad videos (the group "Same-day"), and all videos (the group "All). Topic occurrence ratio was defined as the number of DAT words used in each topic's captions divided by the total number of DAT words used across all topics (so the sum of occurrence ratio over all topics is 100%).

Table 4 displays the alteration rates and the occurrence ratios of different topics in different types of videos. This result provides two important insights about the videos with embedded ads and other videos. First, in videos with embedded ads, two contrasting trends emerged across topics. Highly altered topics, such as Poison/Drug and Police, showed surprisingly lower alteration rates in "Ad" compared to "Same-day", decreasing from 5.41% to 2.56% and from 3.33% to 1.33%, shown in the first and the third rows. However, topics like "Most" and "Blood" experienced much stronger alteration in "Ad" compared to "Same-day", increasing from 2.21% to 3.02% and from 2.91% to 3.40%. Second, the divergences in topic occurrence between videos with ads and those without were relatively minimal across all content categories, shown in the last three columns in Table 4, suggesting that the videos with ads have no substantial topical difference than other videos. This nuanced finding shows that embedded ads did not uniformly reduce alteration rates; instead, they appeared to be associated with higher alterations for platform-warned topics while simultaneously lowering alterations for other topics.

Content	Alteration Rate (%)			Occurrence Ratio(%)		
	Ad	Same-day	All	Ad	Same-day	All
Poison/Drug	2.555	5.408	10.712	3.387	3.507	0.776
Death	2.552	2.558	7.991	17.190	16.082	8.932
Police	1.331	3.332	4.475	8.524	9.193	3.805
Weapon	2.449	5.776	4.187	5.727	5.505	6.363
Kill	3.614	6.547	2.898	13.376	12.388	8.757
Crime	1.855	2.280	2.884	8.043	7.807	4.802
Sex	2.642	5.566	2.827	1.582	1.625	3.725
Blood	3.398	2.907	2.451	4.304	4.015	2.133
State	1.823	1.615	1.797	2.292	2.710	7.457
Most	3.020	2.210	1.787	14.328	15.180	17.035
Others	0.643	1.379	1.773	6.500	7.195	5.310
Luxury	4.354	4.923	1.749	2.125	2.074	1.949
Medical	0.958	0.914	1.624	2.492	2.235	2.482
Ghost	1.456	2.235	1.581	3.280	3.133	2.756
Money	0.915	1.786	1.166	6.849	7.349	23.721

Table 4: Alteration Rate and Occurrence Ratio by Content Topics in Three Types of Videos: alteration rate is calculated by how many of used DAT terms were in altered forms, and occurrence ratio is defined as the use of DAT terms from each topic out of all used DAT terms. "Ad" refers to the videos with embedded ads, "Same-day" refers to the videos without embedded ads but posted on the same day of ad videos from the same channels, and "All" refers to all videos.

In summary, embedded ads are associated with selective and topic-dependent patterns of algospeak rather than a uniform increase or decrease in word alteration. This pattern is broadly consistent with our initial intuition about creator motivation, but points to an unexpected direction: monetized creators appear to respond to perceived moderation risks in a more nuanced and strategic way, rather than by broadly restricting their word usage.

Discussion

In this section, we first situate our findings in the context of Chinese online space and discuss how different factors impact the Chinese users' perception and their algospeak behaviors. Next, we generalize our insights into a more global setting and discuss the relation between algospeak, self-censorship and moderation circumvention as well as the cultural implication of this phenomenon.

Algospeak in the context of Chinese Internet

Our study examines the word alteration practices in a large set of Chinese movie-recap video captions. We find that algospeak is a prevalent phenomenon in the Chinese digital space, but users do not always have a consensus on what terms to alter. This is similar to the qualitative findings that translators having different levels of self-censorship based on their own taken-for-granted in the lack of explicit guidance (Yan 2025). In general, Chinese content creators of movie-recap videos tend to restrain the expressions of negative, violent or absolute meanings. This finding is partially consistent with what Steen, Yurechko, and Klug (2023) and Dawson (2024) described about the TikTok algospeak, where, for example, users would use "unalive" instead of "die". Our dictionary also closely mirrors the YouTube demonetization English vocabulary that users discovered in terms of many topic focuses (Platt 2019).

Next, the user motivation behind algospeak puzzles us because content creators must have a strong reason to "sabotage" their works. We connect the grassroots behavior of video content creators with top-down rules on social media platforms by prompting the two commercial sensitivity detectors. Based on the variation of algospeak preference and proposed reasons from Cizhua and Check51, we argue that the topic alteration tendency in Chinese algospeak was largely resulted from the platform preference, which may be formally written into platform policy or informally demonstrated in moderation actions, and the advertising law in China (King, Keohane, and Verba 1994). Enacted in 2015 and revised in 2018 and 2021, this advertising law explicitly prohibits absolute claims on the Internet such as "the best" or "the first". Because numerous online merchants have been penalized under this law, many Chinese platforms have codified it into their policies and then enforced it strictly through their moderation processes (Douyin E-Commerce 2025). Both the flag reasons from two commercial tools and the frequent alteration in the "Most" terms suggest that the enforcement of the advertising enforcement is an important reason driving Chinese content creators to modify their expressions, which often extends well beyond commercial contexts.

In the meantime, it is noteworthy that our dictionary rarely contains explicit forbidden terms like the variations of Xi Jinping or culturally sensitive topics like LGBTQ. We do not find clear evidence of minority-targeting algospeak examples like what users perceived about the TikTok moderation algorithm (Steen, Yurechko, and Klug 2023; Dawson 2024). We assume this fact reflects the inherent selection bias from content creators: they may already anticipate severe censorship around these topics and the risk of account penalties or audience loss could outweighs the potential benefits of more engagements. As a result of censorship preempting algospeak, politically or culturally sensitive topics may be avoided altogether by content creators and thus be systematically underrepresented in our corpus of movie-recap videos.

Next, we find that monetization through embedded ads in the videos was negatively associated with the alteration rates in the video captions, and there was no evidence that the embedded ads significantly changed the alteration rates in other videos of corresponding channels. However, we do notice that videos with embedded ads usually had a higher alteration rate for the keywords in the "Blood" and "Most" categories. These categories represent the contents that explicitly warned by the platforms like Douyin and the advertising law. These nuanced findings suggest that content creators indeed increased their alteration, though not for all topics, in the videos with commercial interests to prevent being moderated by the platforms. Nevertheless, it remains unclear why content creators relaxed their restrictions on other commonly-altered topics, which ultimately reduced the overall alteration rates for videos with embedded ads. The low correlation between the categories like Most and Police (shown in Appendix) may indicate that users might have different understandings on these topics.

While our videos were collected from YouTube, the prominent existence and high popularity of these content creators on Douyin, as well as their simplified Chinese captions, suggest that their algospeak could be more associated with domestic social media in China instead of other platforms. We are unaware of whether YouTube has such algorithm repression in its recommendation or not, or cannot verify which Chinese video platform has incorporated such algorithmic moderation. However, when creators find content customization too demanding and upload the same video to multiple platforms, the strictest platform exerts the strongest influence on their algospeak behaviors. This phenomenon of cross-posting suggests that creators could unintentionally export the moderation, or censorship, across the platform or the border. In this sense, platform norms, local policies, or laws of dominant groups may extend beyond their original jurisdictions through the cross-regional and cross-platform circulation of content in the online space, inadvertently shaping information environments far beyond their intended scope. (Interestingly, the movies recapped were often not licensed in the mainland China, highlighting an import dynamic from another direction.) This cross-platform content creation behavior also complicates questions of attribution when trying to determine which platform's moderation practices primarily shape creators' behaviors.

Furthermore, Douyin officials repeatedly said that many frequently-altered terms are actually allowed on their platforms and claimed to have the ability to moderate sensitive words even altered. While the Douyin's instruction to avoid absolute expressions is specifically designed for merchants on their platforms, content creators who created movie-recap videos still choose to abide these rules. Objectively speaking, content creators may be convinced by these platform statements to some degree, so they choose to increase their alteration on the clearly-warned topics like medical information and absolute terms in the videos with commercial contents. But their alterations in non-commercial contents indicate that users still anticipate the algorithmic moderation of these platforms will go beyond just the adverting context and target the seemingly harmless contents. This fact of non-compliance suggests that platform does not earn users' trust enough on the matter of moderation (Lin 2025). Because creators observed what happened to others who got penalized for the terms violating the perceived norms or the advertising law, or themselves experienced unpredictable fluctuations in viewership. Lack confidence in the precision of algorithmic moderation, users will rely on their own belief and adopt algospeak in their works.

Therefore, we argue that what matters in algospeak is not what platforms really moderate but what users perceived about the bias or the preference in the moderation algorithm. Algospeak reflects that users reclaim agency over their expressions and restore a sense of order, even illusional, in a complex system that they do not fully comprehend. As platform algorithms grew more opaque, users inevitably resorted to folk theories or seemingly plausible suggestions—much like people turn to religion or supernatural beliefs when confronted with deep uncertainty (Wallace 2013; Savolainen 2022). Amid the trust gap between the platform and the users, altering the words in the video captions may be annoying, but it at least reassures the content creators that they can evade the algorithmic moderation and take control of their works again.

(Chinese) Algospeak as a new Newspeak

Algospeak is a complicated behavior sits between the self-censorship and moderation evasion. While altering the text is an individual action with strong heterogeneity, content creators and their audience implicitly created a codebook of euphemistic terms together, which our dictionary partially revealed. As a collective cultural development on the Internet, algospeak reflects a novel and shared understanding in response to a restrictive environment. On the one hand, this practice of altering the text allows individuals to still communicate their key messages or talk about a topic that is formally or informally prohibited. In contrast to the self-censorship of "not say something", algospeak aligns closer to moderation evasion or censorship circumvention that "manages to say something".

On the other hand, algospeak essentially reinforces a norm where users (voluntarily) conform to imposed standards of content restriction. Compared to the real censorship circumvention behaviors (Hobbs and Roberts 2018; Chang et al. 2022), sidestepping the algorithmic moderation is as much a sign of safer and subtler resistance as it is a sign of constrained user autonomy and effective control, or dancing with chains. Even though the content restriction is believed to be imposed from above, how users adapt contributes to the overall environment of restricted expression. Over time, when a majority of users adopt these alterations of terms in their expressions on social media, it could lead to a self-reinforcing norm where the community collectively shapes what is considered acceptable and then develops a collective unconsciousness about how to express on the Internet. Individual content creators and third party sensitivity detection tools, knowingly or not, proactive or not, internalized the red lines to avoid troubles and participated in the broader information manipulation ecosystem that mimics centralized censorship in its outcomes. Accordingly, algospeak is a spontaneous Orwellian Newspeak fueled by the algorithms.

The impact of algospeak is also culturally profound because the mass media or social media has the power to reshape the culture and language (McLuhan 1994; Androutsopoulos 2006; Van Dijck 2013; Lorenz 2022). Even though we only studied the altered expression in video captions, the twisted audio expression is not uncommon on social media. Users complained about the content restrictions on Douyin and how algospeak manipulated the language and is detrimental to the next generation, and such implication is not limited to Chinese cyberspace (Lorenz 2022).

In addition, we acknowledge the selective bias in our data sampling, because Douyin content creators who chose to cross-post their works on YouTube could be inherently distinctive than the creators who did not. We also acknowledge the limitation that our empirical analysis only focuses on Chinese creators of movie-recap videos, and the specific topical patterns of algospeak may vary across cultures, platforms, and content genres. However, we argue that the core mechanisms identified in this study extend beyond this particular platform–genre pairing. Algospeak is neither unique to Chinese creators nor limited to TikTok- or YouTube-like platforms; rather, it represents a broadly observable form of user adaptation that emerges wherever creators want to regain their agency under complex and opaque moderation regimes. Our findings therefore speak to general dynamics of how creators learn, internalize, and strategically respond to perceived moderation on the platform, even if the surface forms of algospeak differ across contexts. At the same time, we do acknowledge that platforms and content genres differ in affordances, audience expectations, economic incentives, and moderation policies, all of which can affect how algospeak happen in practice. Future work should extend our analysis to other genres and platforms to further understand how users adapt to complicated digital spaces with algospeak or other strategical behaviors.

Taken together, our findings have several implications for digital governance and online ecosystems. First, the topic-selective and incentive-dependent nature of algospeak suggests that content creators are not passive recipients of moderation, but would instead strategically act in response to perceived norms for their own interests. For platform design, this fact brings a significant insight: opaque, strict, or even algorithmic moderation that users do not understand may inadvertently encourage users to adopt distorted expressions even in benign domains, and then impede user experience on the platform. Greater transparency in moderation policies, along with better calibration of moderation algorithms, should help restore user trust, reduce unnecessary algospeak behaviors, and mitigate the decrease of content quality. Second, our results highlight a collective dynamic in which widespread algospeak use reinforces restrictive norms on the platform, even in the absence of direct enforcement. This phenomenon is also important for platform designers who want to foster healthier digital spaces because it suggests that moderation practices shape not only what content is removed, but how communities learn to speak. By empirically examining these user dynamics, our study provides evidence that content moderation should be evaluated not only by accuracy and effectiveness, but also on their longer-term impacts on social norms, trust, and creation.

Conclusion

Our work contributes to the empirical knowledge about user behaviors under algorithmic moderation on social media platforms. We illustrate how they evaded their perceived moderation and what reasons drive their actions. Nevertheless, our study is limited in the scenario of Chinese movie-recap videos and only examines one type of algospeak. Future works could go beyond and examine more algospeak behaviors in different scenarios. Moreover, it is important to investigate the implication of algospeak and algorithmic moderation on our culture outside of the social media world.

In summary, we think algospeak is a concerning phenomenon online. We do not think algospeak could be prevented by educating the users about language usage norms, especially when the governmental restrictions on expressions are still enforced. However, we do believe that narrowing the trust gap between the users and the platform could be beneficial. Fairly speaking, Douyin actually attempts to regain user trust by being more transparent about their algorithm and work process [4]. We would like to see more platforms taking up such responsibilities to build a more trustworthy ecosystem of information. We believe our expressions should be free from both top-down and bottom-up manipulation, and a responsible and reliable platform is essential to this goal.

References

Abokhodair, N.; Skop, Y.; Rüller, S.; Aal, K.; and Elmimouni, H. 2024. Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis. *First Monday*.

Androutsopoulos, J. 2006. Introduction: Sociolinguistics and computer-mediated communication. *Journal of sociolinguistics*, 10(4): 419–438.

Baer, B. J. 2011. Translating queer texts in Soviet Russia: A case study in productive censorship. *Translation Studies*, 4(1): 21–40.

Bao, Y. 2025. ”Calling ’Money’ ’Mi’: Why Is Coded Language Rampant on Platforms Like Douyin?”.

Brunk, J.; Mattern, J.; and Riehle, D. M. 2019. Effect of transparency and trust on acceptance of automatic online comment moderation systems. In *2019 IEEE 21st Conference on Business Informatics (CBI)*, volume 1, 429–435. IEEE.

Calhoun, K.; and Fawcett, A. 2023. “They Edited Out her Nip Nops”: Linguistic Innovation as Textual Censorship Avoidance on TikTok. *Language@ Internet*, 21: 1–30.

[4]https://95152.douyin.com/

Chancellor, S.; Pater, J. A.; Clear, T.; Gilbert, E.; and De Choudhury, M. 2016. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, 1201–1213.

Chang, K.-C.; Hobbs, W. R.; Roberts, M. E.; and Steinert-Threlkeld, Z. C. 2022. COVID-19 increased censorship circumvention and access to sensitive topics in China. *Proceedings of the National Academy of Sciences*, 119(4): e2102818119.

Chen, X.; Xie, J.; Wang, Z.; Shen, B.; and Zhou, Z. 2023. How we express ourselves freely: Censorship, self-censorship, and anti-censorship on a Chinese social media. In *International Conference on Information*, 93–108. Springer.

China Media Project. 2021. Short Take on China’s Short Video Rules.

Cotter, K. 2023. “Shadowbanning is not a thing”: Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*, 26(6): 1226–1243.

Dawson, S. 2024. *You can’t say that on TikTok: cxnsxrshxp, algorithmic (in) visibility, and the threat of representation*. Ph.D. thesis, University of British Columbia.

DeVito, M. A.; Birnholtz, J.; Hancock, J. T.; French, M.; and Liu, S. 2018. How people form folk theories of social media feeds and what it means for how we study self-presentation. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 1–12.

Douyin Blackboard. 2022. Explanation regarding inaccurate moderation rules online.

Douyin E-Commerce. 2025. Douyin E-commerce Creator Standard Terminology Glossary (Version 1.0).

Eslami, M.; Karahalios, K.; Sandvig, C.; Vaccaro, K.; Rickman, A.; Hamilton, K.; and Kirlik, A. 2016. First I” like” it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*, 2371–2382.

Felaco, C. 2025. Making Sense of Algorithm: Exploring TikTok Users’ Awareness of Content Recommendation and Moderation Algorithms. *International Journal of Communication*, 19: 22.

Felaco, C.; Pelliccia, F.; et al. 2024. TikTok algorithm awareness. The case of algospeak to bypass algorithmic logic. In *The Fields of Digital Research: Teoretical, Methodological and Application Challenges*, 66–93. McGraw-Hill Education.

Fillies, J.; and Paschke, A. 2024. Simple LLM based Approach to Counter Algospeak. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, 136–145. Mexico City, Mexico: Association for Computational Linguistics.

Fu, K.-w.; Chan, C.-h.; and Chau, M. 2013. Assessing censorship on microblogs in China: Discriminatory keyword analysis and the real-name registration policy. *IEEE internet computing*, 17(3): 42–50.

Gomez, J. F.; Machado, C.; Paes, L. M.; and Calmon, F. 2024. Algorithmic arbitrariness in content moderation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2234–2253.

Gorwa, R.; Binns, R.; and Katzenbach, C. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1): 2053951719897945.

Guinaudeau, B.; Munger, K.; and Votta, F. 2022. Fifteen seconds of fame: TikTok and the supply side of social video. *Computational Communication Research*, 4(2): 463–485.

Hobbs, W. R.; and Roberts, M. E. 2018. How sudden censorship can increase access to information. *American Political Science Review*, 112(3): 621–636.

Hödl, T.; and Myrach, T. 2023. Content creators between platform control and user autonomy: the role of algorithms and revenue sharing. *Business & Information Systems Engineering*, 65(5): 497–519.

Kim, S. A. 2017. Social media algorithms: Why you see what you see. *Geo. L. Tech. Rev.*, 2: 147.

King, G.; Keohane, R. O.; and Verba, S. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press.

Klug, D.; Steen, E.; and Yurechko, K. 2023. How algorithm awareness impacts algospeak use on TikTok. In *Companion Proceedings of the ACM Web Conference 2023*, 234–237.

Lee, S.-y. 2016. Surviving online censorship in China: Three satirical tactics and their impact. *The China Quarterly*, 228: 1061–1080.

Li, L.; and Zhou, K. 2024. When content moderation is not about content: How Chinese social media platforms moderate content and why it matters. *New Media & Society*, 14614448241263933.

Lin, H. 2025. Oscillation Between Resist and to Not? Users' Folk Theories and Resistance to Algorithmic Curation on Douyin. *Social Media+ Society*, 11(1): 20563051251313610.

Lorenz, T. 2022. Internet 'algospeak' is changing our language in real time, from 'nip nops' to 'le dollar bean'. *The Washington Post*.

Ma, R.; You, Y.; Gui, X.; and Kou, Y. 2023. How do users experience moderation?: A systematic literature review. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2): 1–30.

Mayworm, S.; DeVito, M. A.; Delmonaco, D.; Thach, H.; and Haimson, O. L. 2024. Content moderation folk theories and perceptions of platform spirit among marginalized social media users. *ACM Transactions on Social Computing*, 7(1-4): 1–27.

McLuhan, M. 1994. *Understanding media: The extensions of man*. MIT press.

Meng, Z.; and Nansen, B. 2022. Chinese Video Creator Identities-a Cross-Platform Social Media Perspective. *PLATFORM: Journal of Media & Communication*, 9(1).

Molina, M. D.; and Sundar, S. S. 2022. When AI moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, 27(4): zmac010.

Naher, J.; An, T.; and Kim, J. 2019. Improving users' algorithmic understandability and trust in content moderation. In *The 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing. Association for Computing Machinery*.

Platt, A. 2019. List of YouTube Demonetized Words REVEALED.

Roberts, S. T. 2019. *Behind the screen*. Yale University Press.

Savolainen, L. 2022. The shadow banning controversy: perceived governance and algorithmic folklore. *Media, culture & society*, 44(6): 1091–1109.

Steen, E.; Yurechko, K.; and Klug, D. 2023. You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media+ Society*, 9(3): 20563051231194586.

Suzor, N. P.; West, S. M.; Quodling, A.; and York, J. 2019. What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, 13: 18.

Vaccaro, K.; Sandvig, C.; and Karahalios, K. 2020. "At the End of the Day Facebook Does What It Wants" How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on human-computer interaction*, 4(CSCW2): 1–22.

Van Dijck, J. 2013. *The culture of connectivity: A critical history of social media*. Oxford University Press.

Wallace, A. 2013. *Religion: An anthropological view*. Random House.

Wang, D. 2020. Censorship and manipulation in audiovisual translation. *The Palgrave handbook of audiovisual translation and media accessibility*, 621–643.

Yan, L. 2025. The implications of new censorship theory: Conformity and resistance of subtitle translators in China. *The Journal of Specialised Translation*, (43): 171–187.

Ye, Z.; Huang, Q.; and Krijnen, T. 2025. Douyin's playful platform governance: Platform's self-regulation and content creators' participatory surveillance. *International Journal of Cultural Studies*, 28(1): 80–98.

Zeng, J.; and Kaye, D. B. V. 2022. From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, 14(1): 79–95.

Zhao, A.; and Hu, H. 2025. Unveiling Strategic Governance and User Dynamics in Weibo's Community-driven Content Moderation System. *Journal of Quantitative Description: Digital Media*, 5.

Ethical Checklist

1. Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures?
Yes, and the study advances understanding of user strategies around algorithmic moderation, particularly in Chinese online spaces, without collecting any private user data. We focus on aggregate patterns of text alterations.
2. Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope?
Yes, and the abstract clearly summarizes our main findings.
3. Do you clarify how the proposed methodological approach is appropriate for the claims made?
Yes, and we clearly describe and justify our methods in Section "Methods and Data", linking them to the research questions.
4. Do you clarify what are possible artifacts in the data used, given population-specific distributions?
Yes, and we acknowledge that movie-recap content and language usage are shaped by Chinese social media culture and moderation expectations.
5. Did you describe the limitations of your work?
Yes, as we did in our Discussion section.
6. Did you discuss any potential negative societal impacts of your work?
No, because there is no visible negative societal impact of this work.
7. Did you discuss any potential misuse of your work?
No, because there is no misuse cases of this work.
8. Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings?
NA
9. Have you read the ethics review guidelines and ensured that your paper conforms to them?
Yes, and we adhered to the ethical guidelines for data collection, anonymization, and impact awareness throughout the project.
10. Did you clearly state the assumptions underlying all theoretical results?
NA, our study is primarily empirical and descriptive rather than based on formal theoretical modeling.
11. Have you provided justifications for all theoretical results?
NA
12. Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results?
Yes, and we consider both compliance and strategic behavior theories in interpreting users' alteration decisions in Section 5.1.
13. Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study?
Yes, and we discuss that both platform moderation and user monetization strategies may jointly influence alteration behavior. See Section 5.2.
14. Did you address potential biases or limitations in your theoretical framework?
NA, no formal theoretical framework is presented.
15. Have you related your theoretical results to the existing literature in social science?
Yes, and we reference prior work on content moderation, algospeak, and Chinese internet governance in both Related Work and Discussion sections.
16. Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain?
Yes, and we discuss how findings may inform future platform transparency and user education policies in Discussion.
17. If your work uses existing assets, did you cite the creators?
Yes, and we credited the Github repo and the commercial moderation APIs and video platforms in the Methods and Data section.
18. Did you mention the license of the assets?
Yes, and all video captions were scraped from publicly available content under fair use; commercial APIs are used under their terms of service.
19. Did you include any new assets in the supplemental material or as a URL?
No, because the video dataset includes platform- and channel-level information that could potentially be deanonymized.
20. Did you discuss whether and how consent was obtained from people whose data you're using/curating?
Yes, and the data consists of public content posted by creators on YouTube, which is accessible without login or subscription. We do not analyze user comments or profile data.
21. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
No, but we verify that our dataset does not contain any PII.
22. If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR?
NA
23. If you are curating or releasing new datasets, did you create a Datasheet for the Dataset?
NA

An example of Algospeak

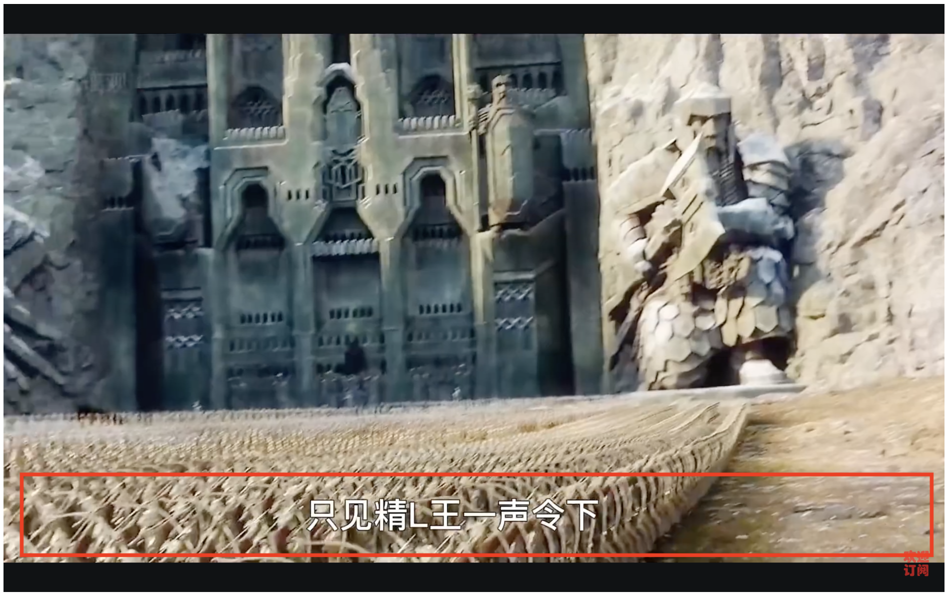


Figure 2: This is a screenshot taken from a movie-recap video about The Hobbit. The word 精灵 (“elf”), which pronounced as Jing Ling in the caption was altered into 精L. This term is likely to be altered because of its superstitious nature instead of other positive or negative meaning. The red rectangle approximately represents the area we searched for and extracted the hard-coded caption text.

An example of Embedded Ad

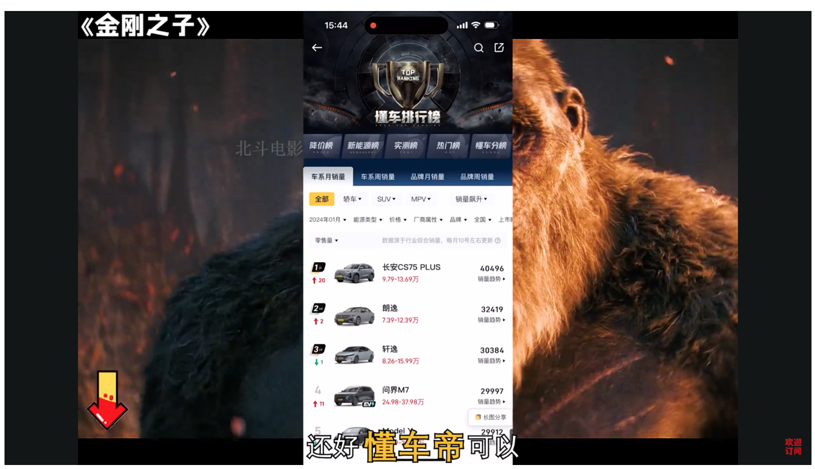
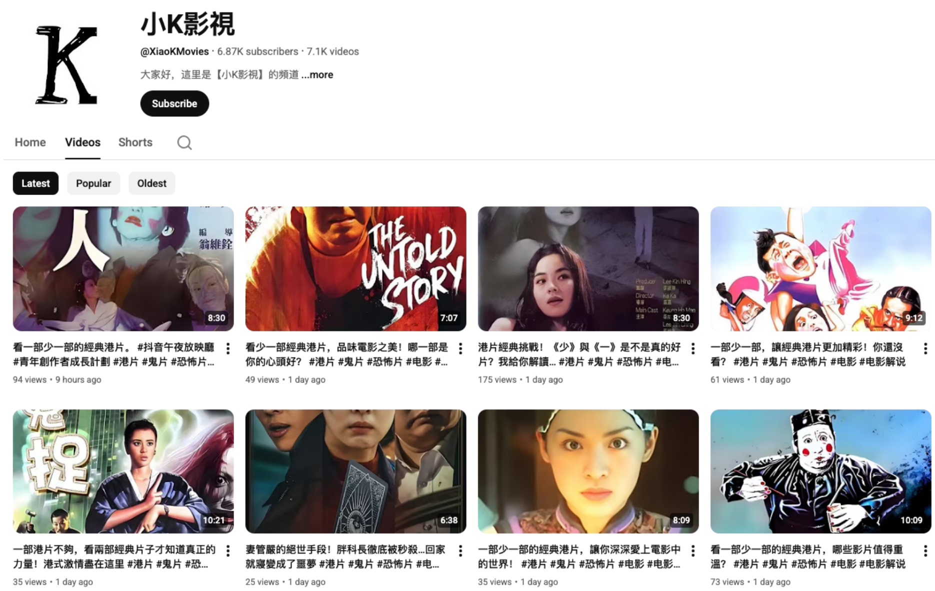
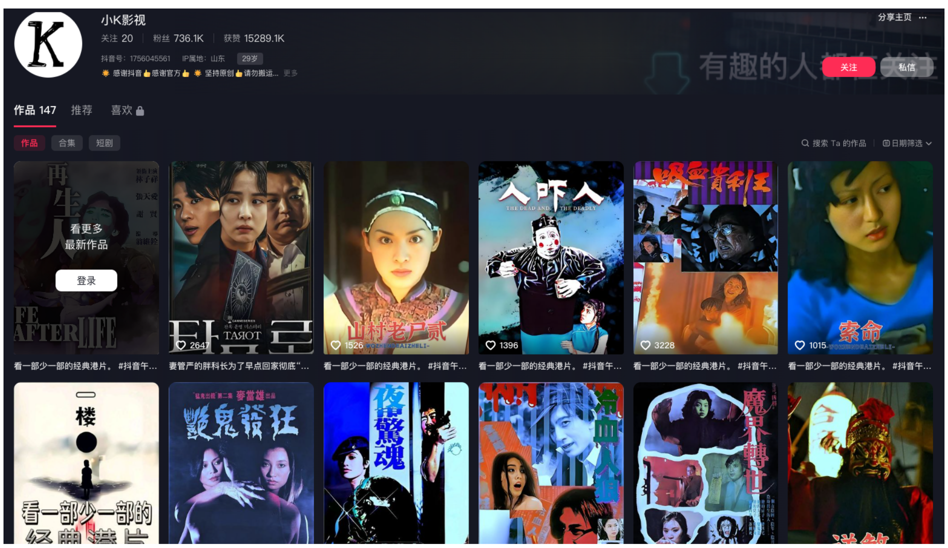


Figure 3: This is a screenshot taken from a movie-recap video with an embedded advertisement. The creator inserted a clip of advertisement for a vehicle information platform. Not only the video frame is covered partly by this ad, the narrative caption is also replaced by this ad.

An example of cross-posting



(a) YouTube Account



(b) Douyin (TikTok) Account

Figure 4: This is an example of one content creator cross-posting contents on both YouTube and Douyin (TikTok).

Prompts to build DAT

System Content

You are a helpful assistant that can help me detect the abnormal words in Chinese subtitles and infer the censored terms. You should find the abnormal terms in Chinese subtitles. These terms usually have one or multiple English letters, which are the initials of Chinese Pinyin of the original Chinese characters. For example, '闹g' is the censored term of '闹鬼', because 鬼has pinyin as 'gui' and the initial is 'g'. You should identify these terms and then infer the original Chinese terms. You should output the all detected results of a sentence in a dictionary, with the abnormal terms as the keys and the inferred terms as the values. Each abnormal term has the minimal length as 2, at least one English letter, and only one correct inferred term. If there is no abnormal term in the text or you cannot confidently infer the original term, you should respond an empty dictionary as "{}". The provided subtitle text could be messy, but you just need to look for the abnormal terms that have Pinyin initials and could ignore English words or common Chinese-English mixed words.

Here are some examples:

Text: 就在他绝望到决定z杀时 Outcome: {‘z杀’:‘自杀’}

Text: 梅川裙子体内检验出了y染色体 Outcome: {}

Text: 蛾男立马发起s频邀约 Outcome: {‘s频’:‘视频’}

Text: ncarbbs这时小帅dna一动，突然想到一个解决办法 Outcome: {}

Text: 上级决定即刻对文姬奶奶执行sx Outcome: {‘sx’:‘死刑’}

Text: 原来秀红的前男友因为偷s漏s被j察抓了 Outcome: {‘偷s’:‘偷税’, ‘漏s’:‘漏税’, ‘j察’:‘警察’}

Text: 年轻的狱警得知女孩怀孕后 Outcome: {}

Text: 724lnb13他准备下一盘大棋而j匪听了这话火上心头 Outcome: {‘j匪’:‘劫匪’}

Text: 田毅却被p了s刑 Outcome: {‘p了:‘判了’, ‘s刑’:‘死刑’}

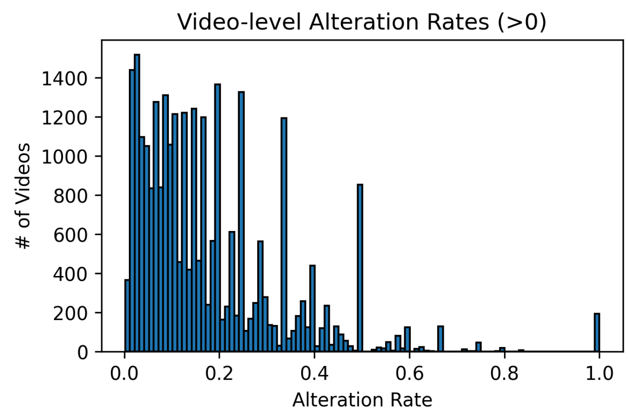
Now, help me analyze the provided subtitle text.

Training Data

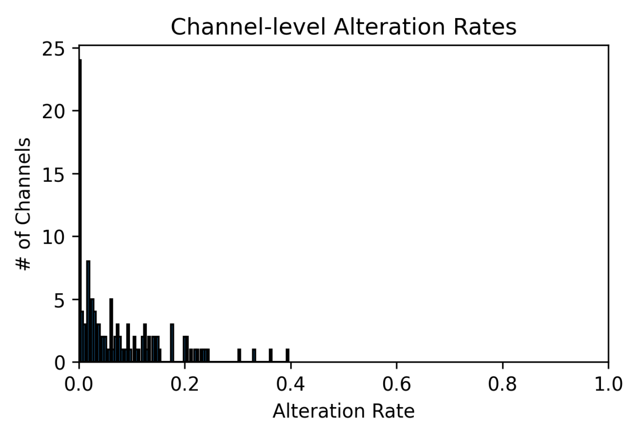
1. **User:** Text: 不死心的王二经常半夜来s扰云秀
   Assistant: Outcome: {‘s扰’:‘骚扰’}
2. **User:** Text: 最终决定将小樱的5年有期tx改为无期
   Assistant: Outcome: {‘tx’:‘徒刑’}
3. **User:** Text: 而此时旁边的养老院正在开party
   Assistant: Outcome: {}
4. **User:** Text: 蟠然醒悟的男人顶着雨拼m往家里跑
   Assistant: Outcome: {‘拼m’:‘拼命’}
5. **User:** Text: hmu又举了很多读写障碍患者的成功案例
   Assistant: Outcome: {}
6. **User:** Text: 大山因为b力违f
   Assistant: Outcome: {‘b力’:‘暴力’, ‘违f’:‘违法’}
7. **User:** Text: 播放包经破千w
   Assistant: Outcome: {‘千w’:‘千万’}
8. **User:** Text: 申然就以工作为由把安琪约到了4s店
   Assistant: Outcome: {}
9. **User:** Text: 拿着ak就是一顿乱扫
   Assistant: Outcome: {}
10. **User:** Text: 那就是他z爱的爸爸
    Assistant: Outcome: {‘z爱’:‘最爱’}
11. **User:** Text: d一季成本就高达4.65亿m元
    Assistant: Outcome: {‘d一季’:‘第一季’, ‘m元’:‘美元’}
12. **User:** Text: 这晚上田义妈妈梦见儿子被l棍d死
    Assistant: Outcome: {‘l棍’:‘乱棍’, ‘d死’:‘打死’}
13. **User:** Text: 他们的爱情故事被翻拍成了一部电视剧
    Assistant: Outcome: {}
14. **User:** Text: takealookatyoursurroundings你看你这里的条件
    Assistant: Outcome: {}
15. **User:** Text: 最后乔西狂赚了100多w
    Assistant: Outcome: {‘100多w’:‘100多万’}
16. **User:** Text: 桥西发现等车的人比做h酸的人还多
    Assistant: Outcome: {‘做h酸’:‘做核酸’}
17. **User:** Text: 只会用下半s思考的c生
    Assistant: Outcome: {‘下半s’:‘下半身’,‘c生’:‘畜生’}
18. **User:** Text: 饥e难耐
    Assistant: Outcome: {‘饥e’:‘饥饿’}
19. **User:** Text: 我们都是将死之人何必折腾不如好好等s
    Assistant: Outcome: {‘等s’:‘等死’}
20. **User:** Text: vizercns当别人想在门口涂鸦广g时
    Assistant: Outcome: {‘广g’:‘广告’}

Alteration Distribution

Based on the video and channel level alteration data, Figure 5 shows the distribution of video-level and channel-level alteration rates in 100 bins. Both subfigures have alteration rates as the x-axis and the numbers of videos or channels as the y-axis. The left subfigure shows the distribution of videos across different ranges of video-level alteration rates, and we excluded the videos with alteration rate as 0 (157,343 videos) to make this figure easier to comprehend. The right subfigure shows the distribution of channels across different ranges of channel-level alteration rates.



(a) Video alteration rate (excluding 0)



(b) Channel alteration rate

Figure 5: Video and Channel Alteration Rates

Temporal Changes

We examine the temporal changes of alteration trends in different topics. Figure 6 shows the frequency of term alterations in a few prominent topics over time. The y-axis represents the total number of videos with any term of a certain topic, regardless altered or not, in their captions. The red bar indicates the number of videos with the altered terms in their captions, and the green bar indicates the number of videos with the original terms in their captions. The numbers on the bar represents the alteration rates of this topic on that month. We find that there is no abrupt or consistent change in alteration rates over months in these prominent topics, as shown in the close alteration rate numbers in Figure 6. Given the limitation of video data time range and most channels on YouTube were created in 2024, we could only conservatively conclude that there was no significant temporal change in terms of topic sensitivity and algospeak behaviors during the time period we observed.

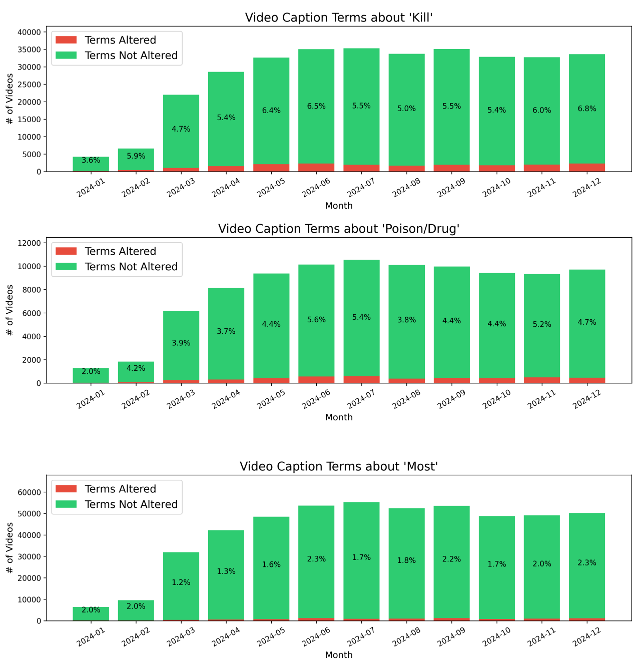


Figure 6: Video Alteration by Topics over time: the percentage numbers in the bars represent how many (%) videos are altered

The Effects of the First Embedded Ad

The surprising results in Table 3 raise a puzzle about the causality between ads and algospeak, as well as the potential cause of this phenomenon. So we chose to investigate whether the first embedded ad of a channel, at least shown in the YouTube videos, has a significant causal impact on the alterations of following video contents. We constructed a binary indicator *Post* as an independent variable to represent whether a date was before (0) or after (1) the moment a certain channel created its first embedded ad. We then estimate two event-study models, one at the daily level (Model 3, covering 15 days before and after the ad) and one at the weekly level (Model 4, covering four weeks before and after). In both models, the dependent variable is the channel-level alteration rate (daily or weekly). We control for channel popularity via the log-transformed total followers and likes, and we include channel fixed effects and time fixed effects (event-day or event-week) to absorb all time-invariant channel characteristics and common shocks. Standard errors were clustered at the channel level to account for within-channel correlation. We expect to see that *Post* variable has a significant coefficient if the first embedded ad somehow changed the subsequent algospeak behaviors of users. However, as Table 5 show, both regressions show that "Post" indicator is not significant in either model ($p > 0.05$), so we have to keep the null hypothesis that the introduction of the embedded ads did not affect channel-level alteration rates.

Statistic	Model 3	Model 4
Dependent Variable	`Daily Alter Rate`	`Weekly Alter Rate`
No. Observations	536	109
Adj. R-squared	0.521	0.822
AIC	-1758.634	-515.050
BIC	-571.929	-445.075
Log-Likelihood	1156.300	283.530
F-statistic	3.508	8.831
Scale	0.00162	0.000423

	Model 3			Model 4		
Variable	**Coef (SE)**	**z**	$P > \|z\|$	**Coef (SE)**	**z**	$P > \|z\|$
Intercept	0.031 (0.002)	12.882	0.000	0.012 (0.005)	2.469	0.014
Post	0.018 (0.017)	1.068	0.286	0.004 (0.007)	0.527	0.598
log_Followers	0.024 (0.001)	26.389	0.000	0.018 (0.001)	16.074	0.000
log_Likes	-0.022 (0.002)	-12.979	0.000	-0.013 (0.001)	-19.169	0.000

Table 5: Regression Results of Two Event Study Models: these tables omit the estimated coefficients for the channel and event-day/week fixed effects for brevity.

Cizhua vs. Check51: Their Differences in General Classification and Reasons for the Top 20 Most Altered Terms

Cizhua and Check51 generally classify flagged terms into four levels: general sensitive, general forbidden, platform-specific sensitive, and platform-specific forbidden (we set the platform parameter as Douyin in our request since YouTube is not an option in these Chinese tools). These two tools often have no consensus on the sensitivity level of the same term. Among all altered terms from content creators, Cizhua labeled 48.7% as general sensitive, 33.1% as platform-specific forbidden, and 16.7% as platform-specific sensitive. In contrast, Check51 categorized 60.8% as general forbidden, 36.8% as general sensitive, and 1.2% as platform-specific sensitive.

Table 6 lists the top twenty most frequently altered words in the movie-recap video captions, along with their associated Cizhua and Check51 censorship reasons. Due to differences in term granularity between our dictionary and the commercial databases (e.g., “Die” vs. “Died”), not every altered term has a directly corresponding reason in Cizhua or Check51. In such cases, we referenced the reason associated with a similar word when possible. Nonetheless, some altered terms still lack any corresponding reason in either Cizhua or Check51.

Table 6: The 20 most altered terms in movie-recap video captions and their associated reasons

Term (En)	Term (CN)	Cizhua Reason	Check51 Reason
Murder	杀人	Uncivilized and violent language	This term may violate advertising law or has appeared in known penalty cases, please make your own judgment.
Suicide	自杀	Uncivilized and violent language	Feudal (reactionary), superstitious, pseudo-scientific words
Killer	杀手	Involving politics, religion, race, gender, terrorism, etc., affecting social stability	This term may violate advertising law or has appeared in known penalty cases, please make your own judgment.
Corpse	尸体	-	-
Death	死亡	Death: Suspected medical terminology	Death: Medical efficacy language (Except for medical, pharmaceutical, and medical device advertisements, any other advertising is prohibited from involving disease treatment functions, and shall not use medical terminology or language that could easily confuse the promoted product with drugs or medical devices.)
Deceased	死去	Death: Suspected medical terminology	Death: Medical efficacy language (Except for medical, pharmaceutical, and medical device advertisements, any other advertising is prohibited from involving disease treatment functions, and shall not use medical terminology or language that could easily confuse the promoted product with drugs or medical devices.)
First	第一	'First' is a superlative adjective used to express the highest degree, typically used in rankings or evaluations. If used to indicate temporal, spatial order, or steps (e.g., "first day"), these phrases do not belong to superlative adjectives expressing degree.	Exaggeration words. When used in advertising, one cannot exaggerate the product's leading position or superiority to avoid misleading consumers.
Murder	杀害	Involves personal safety	This term may violate advertising law or has appeared in known penalty cases, please make your own judgment.

Continued on next page

Table 6 – continued from previous page

Term (En)	Term (CN)	Cizhua Reason	Check51 Reason
Second	第二	It shall not contain content involving ranking, recommendation, designation, determination, selection, award-winning, making a list, spot-checking, testing, statistical analysis, or publication of market survey results that sort or comprehensively evaluate enterprises and their goods or services (except for approved comparative marketing information).	Exaggeration words. When used in advertising, one cannot exaggerate the product's leading position or superiority to avoid misleading consumers.
Police	警方	-	Police Officer: Do not publish police, military uniforms, insignia, equipment, and articles
Last	最后	Initially: If it merely indicates temporal or spatial sequence, or verifiable historical facts that do not change, and if there is factual evidence that can completely and clearly express the meaning without misleading consumers, then it can be used.	Most: absolute, extreme words that cannot be verified. Pay attention to whether combined phrases lack objective standards for measurement, such as 'best' or 'optimal'.
Police Officer	警察	-	Do not publish police, military uniforms, insignia, equipment, and articles
Beat to Death	打死	Beat: Involves violence, endangers personal safety	Beat: This term may violate advertising law or has appeared in known penalty cases, please make your own judgment.
Kill Off	杀掉	Murder: Uncivilized and violent language	Kill: Inappropriate language
Ultimately	最终	Initially: If it merely indicates temporal or spatial sequence, or verifiable historical facts that do not change, and if there is factual evidence that can completely and clearly express the meaning without misleading consumers, then it can be used.	Most: absolute, extreme words that cannot be verified. Pay attention to whether combined phrases lack objective standards for measurement, such as 'best' or 'optimal'.

Continued on next page

Table 6 – continued from previous page

Term (En)	Term (CN)	Cizhua Reason	Check51 Reason
Thoroughly	彻底	Do not include assertive or guarantee language indicating efficacy or safety, such as “safe”, “no side effects”, “non-addictive”, “best therapeutic effect”, “guaranteed cure”, “will cure”, “complete cure”, “immediate effect”.	Absolute, extreme words that cannot be verified
Kill	杀死	Murder: Uncivilized and violent language	Inappropriate language
Handgun	手枪	Rifle: Involves gray industries, illegal industries, illegal tools, prohibited goods, or other content related to illegal activities	Prohibited from publishing information about “firearms, ammunition, military weapons”
Bullet	子弹	Rifle: Involves gray industries, illegal industries, illegal tools, prohibited goods, or other content related to illegal activities	This term may violate advertising law or has appeared in known penalty cases, please make your own judgment.
Must	必须	Extreme words/absolutist language; evidence must be provided to prove they conform to the facts, otherwise they are considered absolutist statements.	Words suspected of being absolute, extreme, and impossible to verify