THE UNIVERSITY OF CHICAGO

# Instagram Shadowbans: Testing for Algorithmic Changes in Treatment of 'Political' Content

By

Grace Shao

June 2025

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Yali Amit
Preceptor: Jon Clindaniel

## Acknowledgements

Many thanks to my cat and roommates for the past year we have shared.

## Abstract

Social media platforms have turned to AI content moderation in response to increased scrutiny for the kind of content they circulate, especially due to major potential for social and political influence. Algorithmic reduction, colloquially known as 'shadowbans,' are favorable for platforms due to their opaqueness, which allows for evasion of responsibility and creates a Foucauldian relationship of platform power and creator discipline. As such, platforms intentionally manufacture opacity surrounding algorithmic reduction. Instagram in particular has recently announced changes in algorithmic content that has caused outrage about their approach to free expression, in part due to vague definitions. This work seeks to shed light on Instagram's intentionally opaque policies. To do so, I define a 'model drift' test to measure change in the ability of a neural network model to predict post like counts. I compare model drift to drift in data in order to tease out algorithmic drift from behavioral drift. I find confirmation that the February 2024 policy change did likely result in algorithmic reduction of content from politicians. More interestingly, I find that pro-Palestine and Zionist content have similar rates of engagement, but only pro-Palestine content shows signs of algorithmic reduction, starting as soon as one week after October 7, 2023. This research serves as a proof of concept demonstrating that intentionally opaque algorithmic reduction can still be investigated. Future research to bring transparency to social media practices is essential for understanding otherwise opaque and guarded practices of algorithmic moderation.

**Keywords:** algorithmic governance, shadowbanning, social media platforms, time series

## 1 Introduction and Literature Review

Meta, the parent company of Facebook and Instagram, has been under heavy public and legal scrutiny to hold the platform accountable for its responsibility as a growing vector of information and social connection. A series of *Wall Street Journal* investigations named "The Facebook Files" revealed how Instagram and Facebook consistently chose to value platform usage and growth more than their negative roles in vaccine information, civic participation, and mental health ("The Facebook Files", 2021). When asked about Facebook's responses to misinformation and hate speech during his 2018 Senate Hearing, CEO Mark

Zuckerberg exclaimed, “AI will fix this!” (Katzenbach, 2021). Indeed, social media companies have turned to artificial intelligence (AI) for content moderation, in large part due to the opacity such algorithms provide. While content moderation is certainly good for hate speech and misinformation, applications in the real world yield concerns of bias and censorship. This project investigates two recent changes in Instagram’s algorithmic content moderation policy to provide some clarity in what topics may have been affected.

The colloquial term ‘shadowbanned’ captures the silent suppression of content: in contrast to the removal of content, where an intervention is out in the light, shadowbanned content leaves no trace. However, many scholars of algorithmic governance find other terms better suited. ‘Shadowbanning’ was first used to describe the power of moderators on early discussion boards to make an offending user invisible to everyone else while maintaining normal functions for the offender (Cole, 2018). Cotter and Gillespie argue that using ‘shadowban’ to describe what modern platforms do today allows them to evade accountability and deny accusations by deflecting outdated definitions and not addressing its colloquial use. ‘Borderline content’ is also commonly used, especially by platforms themselves. However, Gillespie argues it implies a misleading spatial metaphor, as if complex social ideas could be clearly ordered by acceptability. Further, it treats an artificially created line as natural, thus pretending that the policing of such a line is unproblematic (Gillespie, 2022, p. 483). Instead of these terms, Gillespie names the limitation of content circulation through algorithmic recommendation systems as ‘reduction policies,’ the counterpart to algorithmic ‘amplification’ (Gillespie, 2022, pp. 486–492). Accordingly, I use the vocabulary of algorithmic reduction and amplification.

Quantitative research of algorithmic reduction using primary sources (i.e. platform content rather than survey data) is extremely limited. Two studies by King et al. in 2013 and 2014 are seminal in providing insight to what is censored on social media sites in China. While the research methods are very robust, it is not transferable to the context of social media in the modern day and in the west. The researchers in these studies labeled posts as censored or not based on whether the post was taken down or blocked from being published. According to King et al., 2013, 2014, at the time of their study, content filtering was performed by hand and the post was removed if deemed inappropriate. However, since the 2010s, algorithmic recommendation systems have grown in popularity and use. In the 2020s, the Chinese government has worked on implementing regulations of algorithmic recommendation systems to prohibit and promote certain types of content (“China to implement new regulation on algorithm recommendation services”, 2022; “China to tighten regulation of algorithms related to internet information services”, 2021). Unlike the removal of content, there is no confirmation if content has been algorithmically reduced. Additionally, it cannot be assumed that the means and ends of US and Chinese government censorship is the same, or even similar. Another study by Le Merror et al. in 2021 studied shadowbanning on Twitter. However, the study defined shadowbanning if any of three criteria were met: not suggested when searched or mentioned, never shown in search results even if exact username is searched for, and retweets or replies are replaced by “This tweet is unavailable” (Le Merrer et al., 2021). Although this definition of ‘shadowbanning’ shares some characteristics with algorithmic reduction, such as the absence of a notification or confirmation of limited reach, it does not focus on content in algorithmic recommendations, such as a news feed or recommendations page, so I consider it separate from my object of study. Accordingly, the new social media landscape utilizing algorithmic shadowbanning requires a new method and analysis.

## 1.1 History of Algorithmic Reduction

The current landscape of algorithmic reduction in the US can be traced to Section 230 of US telecommunication law. Section 230 provided safe harbor to intermediaries that provided internet access, meaning they were not responsible for what users said, similar to telephone companies and the post office, and would not lose safe harbor if they chose to delete some content. These intermediaries were distinct from content producers, such as radio, television, or video games, that make entertainment as a commodity and were held responsible for the content they produced. Although written decades before social media platforms were popularized, platforms enjoy the protections of Section 230, and take full advantage of the ability to police their content without losing safe harbor (Gillespie, 2018, pp. 28–33, 41).

Before the mid-2010s, platforms worked to position themselves as neutral, in the same category as the original benefactors of Section 230. Companies including Facebook, Twitter, Google, Uber, and AirBnB worked to position themselves as neutral intermediaries, with objective-and-thus-neutral results, and refused to take responsibility for the content that appeared (Ames, 2018; Katzenbach, 2021). While enjoying benefits of Section 230 to the fullest extent, social media companies exist between message delivery and content producers: while they do connect people to people, they also organize content, often through algorithmic selections or recommendations, which effectively function as the commodity consumers pay for with attention and person data (Gillespie, 2018, pp. 34–35, 41).

However, since 2015, more responsibility was placed on platforms, especially as controversies of misinformation and hate speech rose. Platforms adapted to this shift, accepting more responsibility, exemplified by changes in vocabulary and wording from Facebook and Twitter (Katzenbach, 2021). A large part of the attention on platforms as vectors of problematic content is due to unique features that have created unique effects. Compared to traditional media (e.g. newspapers, radio, television) social media relies on user-generated content and has low barriers of entry (Zhuravskaya et al., 2020). Low barriers to entry allows marginalized voices to be heard, but it also spreads extremist opinions and lowers the quality of information.

The Facebook Files demonstrate how Facebook was used to create doubt about the severity of COVID at the beginning of the pandemic and the safety of vaccines ("The Facebook Files", 2021) due to the platforming of non-experts. On the flip side, social media provides a space outside government controlled media, which facilitated the Arab Spring, helping people organize protests and gaining support globally (Smidi & Shahin, 2017). However, authoritarian governments can also use social media to mine online pictures and videos, surveil the cyberspace, and track and arrest protest leaders, as exemplified by Burma in 2007, Iran in 2009, and Bahrain in 2011 (Diamond & Plattner, 2012, pp. xii–xiii). Additionally, governments can also use social media as a tool of propaganda. In 2017, Facebook's algorithmic systems intensified anti-Rohingya sentiment, including content inciting violence and discrimination from top Burmese junta officials, eventually precipitating in the genocide of the Rohingya people ("Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations – new report", 2022). Indeed, the horizontal structure of social media allows for both the weakening and multiplication of propaganda, evasion of censorship and expansion of surveillance (Morozov, 2012, p. 83). It can even be argued that many of these effects can be seen in the US: TikTok increased support for Palestine (Conley, 2024), RedNote showed the lifestyle of Chinese citizens to be different from what many Americans believed (Weedston, 2025), 'fake news' circulated on Twitter heavily contributed to Donald Trump's first electoral win (Allcott & Gentzkow, 2017), and AI bots help police surveil protests on social media (Maiberg & Koebler, 2025).

In response to the shift in responsibility, platforms have turned to AI. In this situation, AI is a 'technological fix,' a technology used as a solution to a nontechnical issue, mixing blind faith in technology with ignorance of root social causes: "traffic management systems cope with the increasing number of cars in cities (and car traffic is not reduced), food imports keep the poorest from starving (and the causes are not combated), cattle are culled (and industrial factory farming is not adopted)" (Delege, 2002, cited and translated in Katzenbach, 2021, p. 2). In the case of social media, algorithms are positioned as a necessary and effective solution: platforms argue that only automated systems are able to handle the scale of content, and so AI solutions appear inevitable. Yet algorithms are only a band-aid to social and legal pressures, leaving social conditions unaddressed. In the 2018 Senate Hearing, Zuckerberg responded to questions about Facebook's plans to handle disinformation campaigns and the spread of hate speech by repeatedly pointing to AI detection systems as a solution (Katzenbach, 2021). In doing so, he avoided talking about how Facebook's structure creates the conditions for such information to spread, such as repeated executive decisions to choose user engagement and platform growth over users' mental health and quality of information ("The Facebook Files", 2021). Thus, rather than address underlying root causes that contribute to the circulation of problematic content, Zuckerberg instead focused on algorithmic filtration systems.

## 1.2 Incentives and Responses to Algorithmic Reduction

Platforms are motivated to use reduction policies rather than removal for several reasons outside of public interests. Reduction is less noticeable, and thus less politically risky than removal, since it evades accusations of bias and censorship. Further, in terms of monetary motivations, reduction allows platforms to continue financially benefiting from users who may otherwise be unhappy with removal, while addressing public concerns by reducing reach. Reduction also allows for obscure policies, giving platforms the flexibility to respond to changing problems (such as content designed for loopholes) but also avoid accountability and clear articulation (Gillespie, 2022, pp. 487–490). Specifically, the use of algorithms to reduce content allows platforms to take advantage of the magical and mystic quality ascribed to algorithms, wherein the impression of algorithms as a black box leaves little hope for supervision or regulation, even though efforts to interrogate algorithms do exist and have progressed (Ames, 2018, p. 2). This further allows platforms to evade responsibility, and places the onus on creators to decipher what sort of content might be affected. The manufactured opaqueness results in a Foucauldian relationship of platform power and creator discipline, where the constant threat of invisibility pushes creators, who often depend on visibility for an income, to self-police their content (Duffy & Meisner, 2023, pp. 300–301). Such opacity creates a power imbalance, where platforms hold the epistemic authority on their algorithms – leading to "black box gaslighting," which makes users doubt themselves, ultimately destabilizing possibly credible accusations (Cotter, 2023, pp. 1226–1227).

Instagram is a large offender of black box gaslighting, using their epistemic authority to discredit accusations of bias through multiple explanations. They have used semantic strategies to deny accusations of shadowbans both by defining shadowbanning as content removal and by explaining specific cases to narrow the issue (e.g. too many hashtags). Instagram authorities point to three main explanations to debunk shadowbans. First, they relabel such incidents as glitches: "Shadow banning does not exist, it is a persistent myth [...] it's because there was legitimately a bug that was affecting hashtags" (Instagram's Director of Fashion Partnerships, quoted in Cotter, 2023, p. 1235). Second, Instagram flips the blame to the creators, explaining that a lack of engagement is simply failure to create engaging content. Third, Instagram points to chance and outside forces: "Even when ranking doesn't change at all, too much else changes in the world. What people are interested in changes, what else you're competing with changes" (CEO Adam Mosseri, quoted in Cotter, 2023, p. 1235).

Due to the prioritization of advertiser appeal, algorithmic systematic policing goes far beyond what is required by law and is more relevant than legal restrictions in terms of public discourse and lived user experience (Gillespie, 2018, pp. 30–35). As a result, many creators have felt their voices were unfairly algorithmically demoted. Since many creators often rely on visibility for an income, there is strong motivation to understand and avoid algorithmic reduction (Nicholas, 2022, p. 31). Due to the lack of clarity from platforms themselves, creators often rely on and formulate theories around algorithmic reduction through gossip with friends, mutually followed users, or on forums. Multiple qualitative studies, using surveys or interviews, have studied the creation and content of these theories. Creators look for signs of algorithmic reduction through drops in engagement (compared to previous average engagement) or followers notifying the creator that creator's posts are not reaching them (Cotter, 2023; Duffy & Meisner, 2023; Nicholas, 2022). Through the framework of institutional power, governance takes place through the process of discipline, where punishments of (in)visibility leads to the strict discipline of creators, who create and follow theories of algorithmic amplification and reduction (Duffy & Meisner, 2023). These theories form as people find patterns in group settings or through experiments comparing what is favored (e.g. blonde versus brown hair, large versus small following) (Duffy & Meisner, 2023). Creators also try to post what they believe the algorithm "likes", such a preference for images posts over posts with links, images of faces, or changing the framing to focus on education and civil liberties (Broniatowski et al., 2020; Horten, 2024).

## 1.3 Instagram and Algorithmic Reduction

Meta has been working on suppressing political content on multiple fronts. Facebook came under fire for its role in the organization of the January 6th insurrection, and company executives were called in for congressional hearings (Horwitz et al., 2023; Telford, 2021). A month later, the company began working on the suppression of political content (Stpanov & Gupta, 2021). While Zuckerberg claimed the limitation of political content comes from community feedback, *The Wall Street Journal* reported that Meta's leaders were turning to algorithmic suppression to evade further criticism as a result of the January 6th attack (Horwitz et al., 2023).

In February 2024, Instagram announced it would no longer "proactively recommend content about politics on recommendation surfaces," defining political content as content "potentially related to things like laws, elections, or social topics" ("Update on Political Content on Instagram and Threads — about.instagram.com", 2024). This announcement has caused wide accusation of censorship of the genocide in Gaza (de Guzman, 2024), given that short form videos (popular on Instagram Reels and TikTok) have had a large role in documenting and spreading information about the genocide. Meta has a history of censoring Palestinian voices: in 2021, when the escalation in violence and forced displacement of Palestinian homes in Sheikh Jarrah (located in occupied East Jerusalem), *Human Rights Watch* found that Meta was suppressing content posted by Palestinians and their supporters ("Israel/Palestine: Facebook Censors Discussion of Rights Issues", 2021). A report by Business for Social Responsibility (BSR) commissioned by Meta found that Meta's actions "appear to have had an adverse human rights impact...on the rights of Palestinian users to freedom of expression, freedom of assembly, political participation, and non-discrimination, and therefore on the ability of Palestinians to share information and insights about their experiences as they occurred" ("Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021", 2022).

TikTok and Instagram have given the public direct, first hand accounts of Israel's attacks, including information about the children, hospitals, and universities. Both Instagram and TikTok have provided a platform for those in Gaza to document their genocide, but TikTok is a Chinese company while Instagram is US owned. Senator Mitt Romney, in an "incredible mask-off moment" admitted the push to ban TikTok is supposed to shut down Americans' access to unfiltered news about the Israeli assault on Gaza (Conley, 2024). Representative Mike Lawler (R-N.Y.), who co-sponsored the bill to ban TikTok pointed to the protests on college campuses as the reason for the TikTok bill (Thakker & Lacy, 2024). The CEO of the Anti-Defamation League (ADL), the largest pro-Israel lobbying group, said in a leaked phone call "We really have a TikTok problem," in the US (Perkins, 2024). The absence of pushes to ban Instagram begs the question whether the platform has changed the access and circulation of information about Gaza to align with US government interests. In response to accusations about the February 2024 policy censoring pro-Palestine content, Meta released a statement that they "apply these policies equally around the world and there is no truth to the suggestion that [they] are deliberately suppressing voice" ("Meta's Ongoing Efforts Regarding the Israel-Hamas War", 2023). In the same statement, Meta announced they fixed bugs that reduced reach, which had "nothing to do with the subject matter of the content," continuing their practice of black box gaslighting by relabeling these instances as glitches.

In an update January 2025, Instagram announced that it would 1) end third party fact-checking in favor of community notes 2) take a "personalized approach" to political content, and 3) allow users to call LGBTQ people mentally ill (Kaplan, 2025; Lavietes, 2025). While the policy does state it would not reduce political content in algorithmic spaces, it does not necessarily mean their algorithmic recommendation system returns to what it was pre-February 2024.

### 1.4 Research Motivations

While existing literature has extensively examined user theories and platform policies, quantitative research on content reduction is sparse. Most existing research on reduction policies are based on interviews, online discourse, or surveys (Cotter, 2023; Duffy & Meisner, 2023; Zeng & Kaye, 2022). Large scale research of social media censorship is sparse, and does not fit definitions of algorithmic reduction (King et al., 2013, 2014; Nicholas, 2022). The circumstances and mechanisms of algorithmic reduction, as seen above, are completely different. The literature emphasizes the importance of need for clarity, and lack thereof, surrounding algorithmic reduction. Platforms have turned to algorithmic solutions when faced with social issues (Katzenbach, 2021). They are actors with their own motivations, and resulting patterns of reduction are biased and prioritize economic gains over user concerns (de Guzman, 2024; Katzenbach, 2021; Telford, 2021). By reducing reach rather than removing content, platforms take advantage of algorithmic obscurity to discredit creator experiences and understandings (Cotter, 2023; Duffy & Meisner, 2023). With incomes, livelihoods, and civic participation on the line (de Guzman, 2024; Duffy & Meisner, 2023; “The Facebook Files”, 2021), transparency – from the platforms themselves or not – is needed. In this project, I look to shed light on Instagram’s algorithmic reduction of political content, and what that may include.

## 2 Data and Methods

### 2.1 Data

The dataset includes 163,179 posts across 312 accounts sorted in fourteen groups. For each post, I collected post time, like count (at time of collection), and caption. The groups analyzed are accounts posted about or owned by: cats, cooking, brands, celebrities, Democrat politicians, Republican politicians, news outlets, right-wing health (e.g. raw milk and anti-vaccine), left-wing health (e.g. disability rights and COVID), LGBTQ (with a focus on transgender and genderqueer accounts), gun rights, tradwives (traditional wives), Pro-Palestine, and Zionist accounts. For a full list of the accounts included, see Appendix A:.1. Categories were chosen by topics close to each other, with varying levels of political relevance. Cats, cooking, brands, and celebrities represent categories expected to be furthest from politics. Of course, nothing is absent of politics: brands and celebrities are tied to capitalism, class, and political associations; cooking is a large part of tradwife content; and conservatives prefer dogs over cats (Ivanski et al., 2021). However, these categories are meant to be less directly political than the others. Democrats, Republicans, and news outlets represent groups of accounts from authority figures, with politicized content. Right-wing and left-wing health compare health trends between different political views. Gun rights, tradwives, and LGBTQ accounts are meant to compare gender stereotypes and flexibility, as gun rights advocates usually buy into masculine narratives and stereotypes, tradwives represent a return to conservative definitions of a wife and femininity, while LGBTQ accounts that focus on gender explore existence outside of a binary understanding of gender.

To choose which accounts would be included, I started with 'parent' accounts in different topics, and collected usernames of accounts the parents followed, or the 'children.' In the end, this preliminary dataset included 10,192 children. For each child, I collected how many followers and posts they have, and the captions of their most recent 12 posts. Since I am investigating algorithmic reduction, I am interested in accounts that regularly appear on algorithmic feeds. The more followers and posts an account has, the more likely they are to be a public facing account aiming to show up on algorithmic feeds and be impacted by algorithmic reduction. Accordingly, I only consider accounts with at least 3,000 followers and 50 posts. To choose a final set of thirty accounts per group for data collection, I used a several methods, depending on the group.

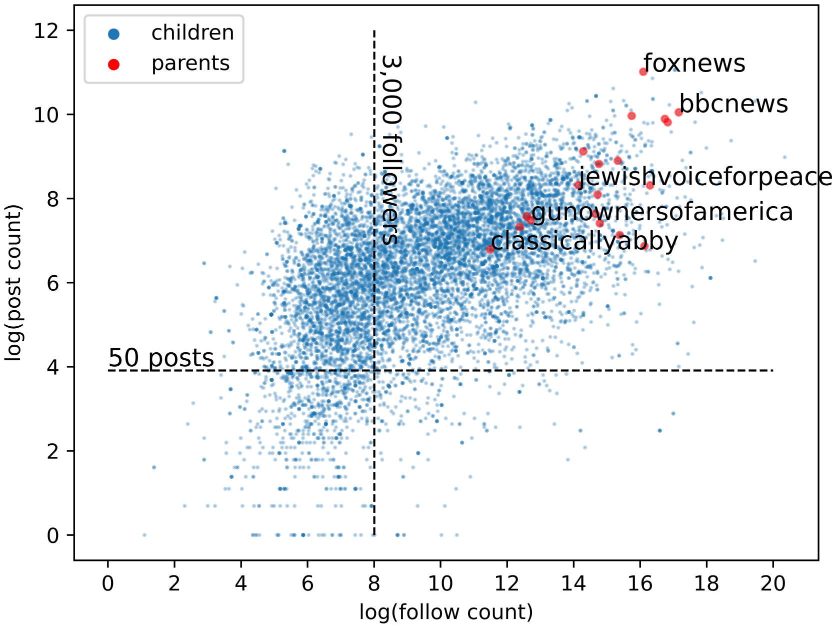

Figure 1: Follow and Post Count of Preliminary Dataset
A plot of log post count versus log follow count. One dot represents one account. All users in the preliminary dataset are included. Parents in red give an idea of where popular accounts in different categories are.

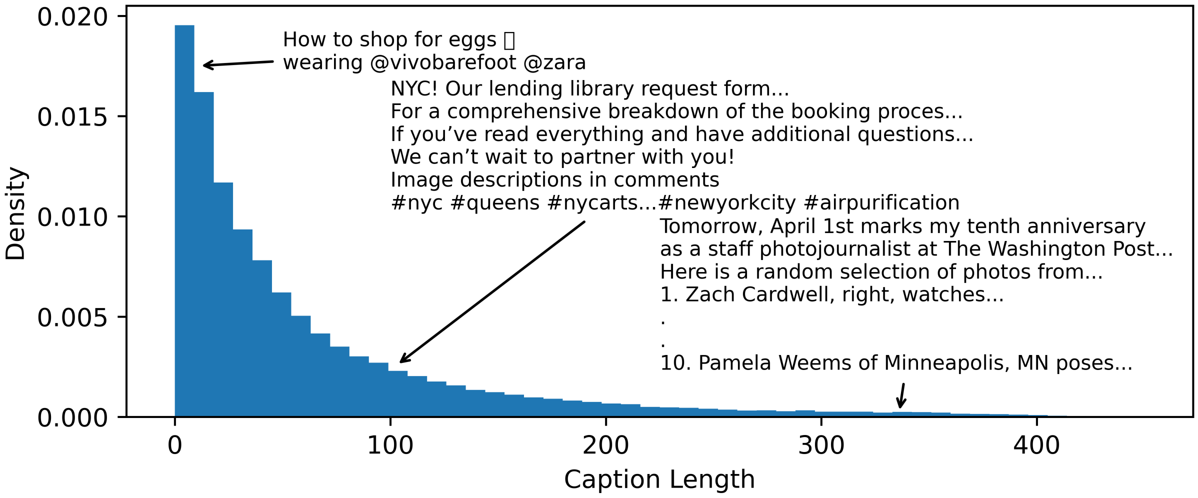

Figure 2: Distribution of caption length
Example captions from @dariasavaya, @a.i.r.nyc, and @mattmcclainphoto, respectively.
Captions are split into tokens by whitespace.

- **LDA Topic Modeling:** Using the captions from the twelve most recent posts, I ran an Latent Dirichlet Allocation (LDA) to model the topics. A document here is defined as the twelve concatenated captions, so one document represents one account. Across n = 5, 10, ..., 50 number of topics, the model with the best coherence score had 35 topics. Since only a few of these topics easily fit into my pre-defined topics, I only used LDA topic modelling to choose accounts in topics that were easily defined. For example, while the "kid-mom-build-easy-sleep" topic may fit the category of tradwives, it likely includes many other accounts. By contrast, "recipe-add-cup-oil-ingredient" easily fits into "cooking."

- **Keyword Search:** Some topics are easily defined by the presence of a word: for example, accounts where "cat" appears in the caption (or username) are likely accounts that post about cats.

- **SBert Embeddings:** For topics that were still semantically similar, but did not coalesce around one or two keywords (such as LGBTQ, where keywords are many), I use SentenceBert (SBERT) (Reimers & Gurevych, 2019) to find accounts with the closest cosine similarity to the average embedding of the parent accounts. Again, a document is twelve concatenated captions, one document for one account. Since SBERT has a context window of 512 tokens, and a small portion of documents exceeded this context window, I chunk longer documents into 512 tokens with a 128 token overlap and average the embedding across the chunks.

- **Snowball Sampling:** Some groups required more nuance, since they post about semantically similar topics but are distinctly separated by opinion. For example, Pro-Palestine and Zionist accounts may use many similar keywords and have similar embeddings, but differ quite strongly in substance. There were only a few accounts in the preliminary dataset belonging to these groups that I could find using the above techniques, so I started with the accounts I could find and snowball sampled from there. I visited each account and found accounts they were following which were popular. I also clicked on hashtags to find other popular accounts posting about the topic.
- **Generative AI Suggestions:** Some topics have accounts that post about many different topics: for example, politicians and news stations are each distinct groups, but the captions are not semantically similar. Thus, I found that my previous work to download children of parents unusable for these groups. Instead, I used a generative model, DeepSeek R1, to find sources. The exact prompt used was "Give me 50 [group description] Instagram usernames (without the @ symbol), formatted as a comma separated list."

All methods described above are used as a preliminary filter. I then sorted all accounts by descending number of followers, and visited each account to check that they were public facing and indeed were focused on the selected topic. I included about thirty accounts for each group, more or less depending on ease of access.

For each account, I collected all information on all posts from January 2023 to March 2025, including caption, like count, and post time. While I attempted all accounts in the final set of usernames, I was not able to successfully collect data for all existing posts within the analysis period for all accounts. I did not include accounts that did not have full data. The news category was disproportionately impacted, likely due to the immense amount of posts news accounts have (with multiple posts per day). One account from the LGBTQ group was removed for being an outlier in post frequency, hovering around four log posts per week, while the rest of the accounts are below two. Additionally, some accounts choose to hide like counts on certain posts, which displays "Liked by [username] and others" instead of "[number] likes" to users. In data collection, these posts are recorded to have -1 likes, so I do not include these posts in the data.

In all my analyses, I use the log likes rather than the raw like count, since the distribution of likes has an extreme right skew, with a median of 3,204 and a mean 174,482. The log likes are more evenly distributed, with a mean of 8.25 and a median of 8.07. This is what I use to analyze algorithmic reduction. There are many ways consumers interact with content

| Group | Attempted Accounts | Accounts | Posts |
|---|---:|---:|---|
| Cat | 31 | 30 | 7,422 |
| Cooking | 35 | 30 | 10,264 |
| Brands | 30 | 30 | 31,802 |
| Celebrities | 37 | 35 | 11,405 |
| Democrats | 31 | 30 | 13,603 |
| Republicans | 31 | 29 | 13,222 |
| News | 31 | 5 | 5,845 |
| Health Right | 14 | 12 | 6,571 |
| Health Left | 17 | 8 | 694 |
| LGBTQ | 20 | 19 | 9,400 |
| Gun | 30 | 30 | 17,260 |
| Tradwives | 11 | 8 | 1,290 |
| Pro-Palestine | 36 | 25 | 11,428 |
| Zionists | 35 | 21 | 22,973 |

Table 1: Number of Accounts and Posts per Group

– views, likes, comments, and sharing. Content views, while available for reels, are not available for posts. I consider likes to be the best measure of algorithmic attention, since each user can only like once (versus multiple comments or shares), and they are the lowest measure of interaction, after a view. To comment or share a post, the post must have made some deeper impression on a user, while the barrier to liking a post is low.

The trends in log likes and post frequency can be seen in Figure 3 and Figure 4. In Figure 3, each row shows pairings of groups in similar topics. The figure plots rolling means of log likes normalized within each group to depict relative interest in groups over time. For example, the clearest trend can be seen in pro-Palestine and Zionist content, where the content in these categories experienced explosive growth right after October 2023. Other trends are also seen here: tradwife content experienced the most engagement relative to itself around spring of 2024. Content from Republican and Democrat politicians grew leading up to the election in November 2024, and falling soon after. Brands experienced a jump right before January 2025. While these plots allow for comparison of normalized log-likes within a group, they cannot compare interest across groups. For example, although Zionist content has higher normalized log-likes than pro-Palestine content before October 2023, it does not mean that Zionist content was more popular than pro-Palestine content. Rather, both experienced the lowest engagement in the dataset before October 2023, and jumped in engagement in October. The groups are plotted together to compare trends that may have happened across groups with similar topics, such as pro-Palestine and Zionist content increasing due to a relevant world event or Republicans and Democrats preparing for the election. In Figure 4, the log of posts per week are plotted by account, sorted by group. This figure demonstrates that while posting frequency fluctuates for accounts, there are no major coinciding changes in the accounts that post the most. Accordingly, the proportion of posts from each account in the training data in the before set should relatively reflect the proportion in the after set. Looking at specific groups, a major jump in post frequency occurs for pro-Palestine and Zionist accounts. The posting frequency remains consistent for Zionist accounts, but not for pro-Palestine accounts, which decrease in frequency, especially top posters.

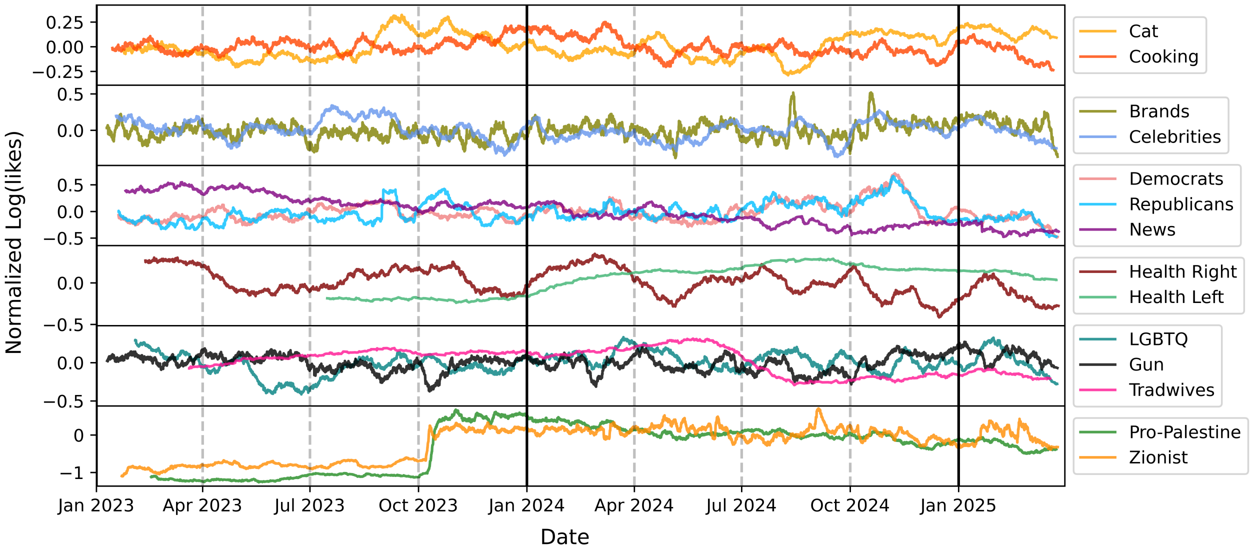

Figure 3: Growth of groups
A plot of normalized log likes by time.

## 2.2 Methods

To test for a change in algorithmic reduction, I sort data by group and split into 'before' and 'after,' where before includes posts assumed to be before any algorithmic testing, and after includes posts after the announcement from Instagram. I train a model on the before data to predict log likes of a post based on features of the current post and previous posts. Using this model, I collect the residuals of the before and after data. To draw out algorithmic drift, I compare before and after with t-tests for the difference in means of residuals and of

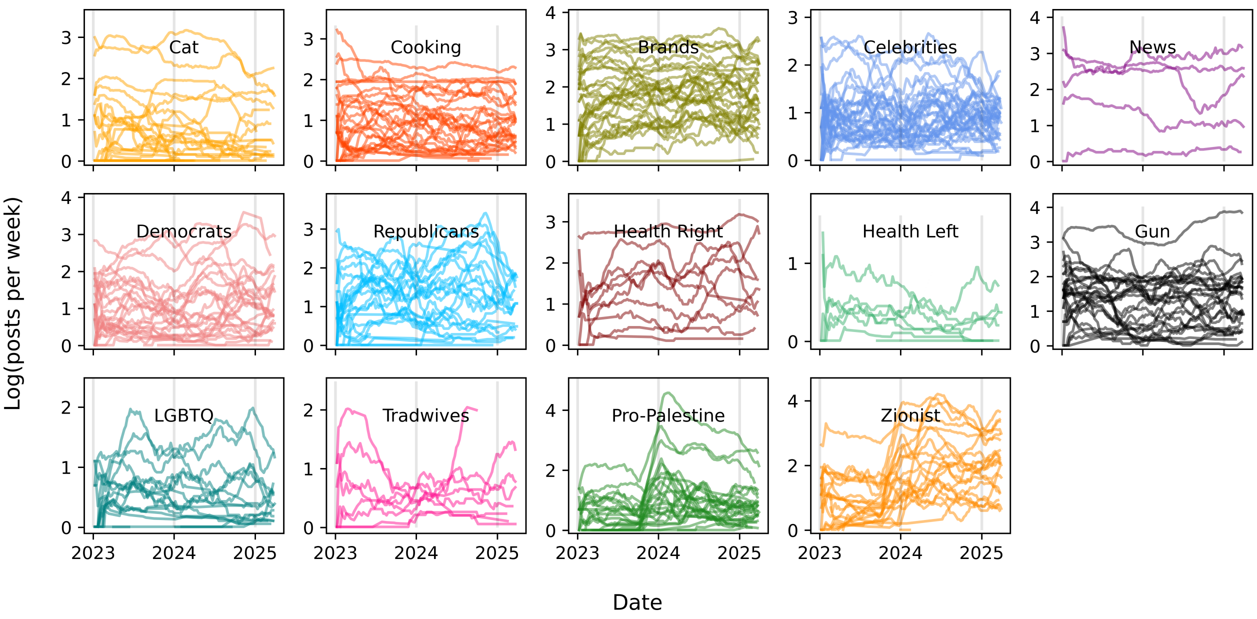

Figure 4: Frequency of posts by account
Plots of log posts per week by time.

log likes. I call the t-test for difference in means of log likes the **test for data drift**, since this test captures how the data itself has changed. I call the t-test for difference in means of model residuals the **test for model drift**, since this test captures how the model has drifted in ability to capture the data. For shorthand, I refer to the t-statistics from these tests as the data drift t-statistic or model drift t-statistic.

A positive data drift t-statistic indicates the mean log likes of the after set is less than the mean log likes of the before set, showing a decrease in the group's popularity. Similarly, a negative data drift t-statistic indicates an increase in the group's popularity. A positive model drift t-statistic indicates the model consistently underestimates the log likes for the data in the after set, while a negative model drift t-statistic indicate the model consistently overestimates log likes for the data in the after set. Patterns in over- and under- estimation show that the data is being treated differently, either by users or the algorithm. These changes are due to behavioral and system drift (Salganik, 2019). Behavioral drift includes changes in user interest and interaction of content, while system drift describes changes in the system itself, such as changes in algorithmic reduction and amplification.

While engagement and algorithmic recommendations are heavily intertwined, as algorithmic recommendation systems rely on engagement metrics, and engagement depends on algorithmic decisions, the interpretation of these two t-statistics are still separate. The difference is most clear in a case where the two disagree. When the signs of the two agree, it's likely that the under/overestimation of the model is due to a difference in mean log likes, and algorithmic changes either add to the effect or are small enough that it does not take away from the change in popularity.

A disagreement in sign is likely a signal of algorithmic reduction or amplification. For example, a category might generally be decreasing in popularity and getting less log likes as time goes on, but is algorithmically amplified for whatever reason (e.g. paid ads, bot accounts). In this case, the data drift t-statistic may be positive, while the model drift t-statistic might be negative. Conversely, a category might be increasing in popularity, but is algorithmically reduced, and may accordingly have a negative data drift t-statistic of log likes with a positive model drift t-statistic on residuals. Significant t-statistics are those with p-values less than 0.05, after adjusting raw p-values using the Bonferroni correction by multiplying by fourteen (the number of categories).

| Name | T-test on | Sign | Interpretation |
|---|---|---|---|
| Test for Data Drift | Log likes | Positive | decreased engagement |
| | | Negative | increased engagement |
| Test for Model Drift | Residuals | Positive | decreased engagement and/or algorithmic reduction |
| | | Negative | increased engagement and/or algorithmic amplification |

Table 2: Interpretation of Tests

## 2.3 Data splits

There are two pairs of events and Instagram policy changes I am interested in for this study. In the first policy change on February 7th, 2024, Instagram announced they would algorithmically reduce political content. I pair this policy change with October 7th, 2023 as the motivating event, when Hamas initiated an attack of resistance on Israel, which became a turning point for the escalation of the occupation and genocide in Gaza. The second policy change on January 7th, 2025, announced that Instagram would 1) end third party fact-checking in favor of community notes 2) take a personalized approach to political content, and 3) allow users to call LGBTQ people mentally ill (Kaplan, 2025; Lavietes, 2025). I pair this policy change with November 6th, 2025 as the motivating event, the date the 2024 presidential election concluded, announcing Trump as the next president.

For both policy changes, the date of the likely motivation (October 7th and November 6th) does not represent the start of their respective issues. Israel's oppression and occupation

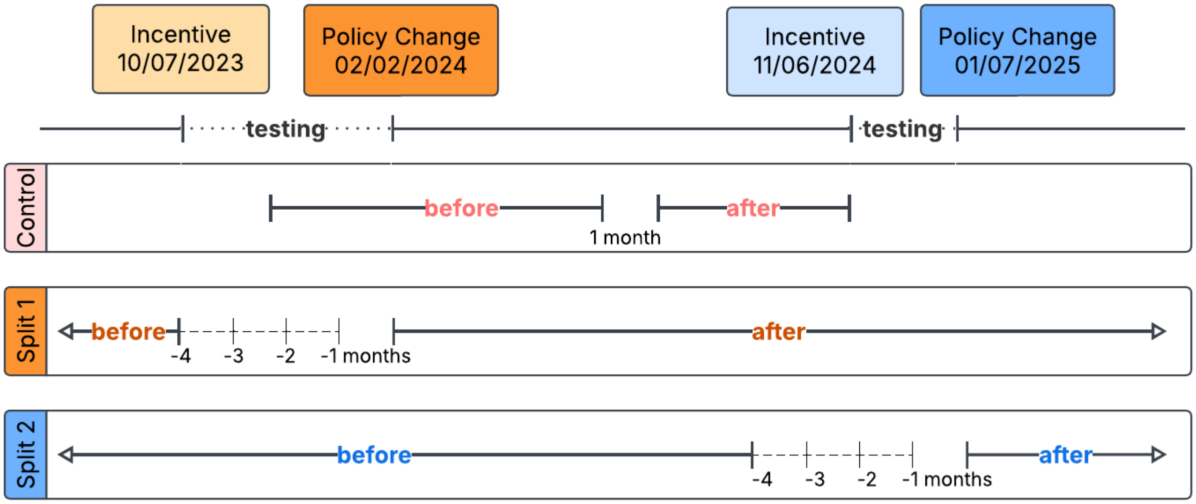

Figure 5: Timeline of relevant events and data splits

of Palestine has a history that extends half a century before October 7th, 2023 ("World Court Finds Israel Responsible for Apartheid", 2024). Similarly, far right ideology in the US did not start when Trump was elected for a second term. These dates do not represent the start of these issues, but instead mark a date of incentive for Instagram's policies to change. After October 7th, 2023, the human rights violations and discrimination of the Palestinian people became more relevant in the US, in part due to short-form videos. The unfiltered documentation of the genocide goes against the US government's interests, leading to movement to ban TikTok, and an incentive for Instagram to change how such information disseminates. After November 6th, 2024, it became clear that Trump would have a second term. Before the election, it was unclear what policies on Instagram would lead to better government relations, especially after a show of force to ban TikTok. The election results served as incentive for Instagram to align itself with the incoming administration.

I use a period of four, three, two, and one month(s) before the announcement to mark the period between before and after, a time when Instagram may have been testing algorithmic changes. The first split uses all data after February 2024 in the 'after' set, and data up to October (-4), November (-3), December 2023 (-2), and January 2024 (-1) is used in four 'before' sets. These before sets have the additional advantage of varying how much data after October 2023, where content on Palestine and Israel jumped in log likes, is included from zero, one, two, and three months of data. The second split uses all data after January 2025 in the 'after' set, and data up to August (-4), October (-3), November (-2), and

December (-1) is used in four ‘before’ sets. Three of these sets do not include any data after Trump’s win, and the last set includes one month of data after Trump’s win.

As a control, I use the data between the first policy announcement and the second event to compare model architectures and scale of t-statistics. This is a period of less volatility, where the log likes of groups are relatively stable (as compared to the jump for Israel/Palestine content around October 2023 or the peak of politicians around November 2024). For the control, data between February 2023 and July 2023 forms the ‘before’ set, and data between August 2023 and November 2023 forms the ‘after’ set.

### 2.3.1 Model Architectures and Features

Since the likes of a post depend on many complicated variables including post time and caption content, I used three architectures of neural networks: one- , two- , and three-branched architecture. I train the models on the before data of each group and pass each of these models to test for model drift. More negative residuals indicate the model consistently overestimates performance of the after set, while more positive residuals indicate the model consistently underestimates performance of the after set. Changes in residual distribution is due to behavioral or system drift. In this case, behavioral drift would be users’ actual behavior of content they engage with: how interest in content changes. Algorithmic changes are system drift: how the content being shown to users changes. The two are heavily intertwined, since changes in what content is reduced or amplified influences what content users are interested in. While the residuals themselves do not differentiate between the two types of drift, comparing model drift to data drift can indicate algorithmic changes. Additionally, information about relevant events, such as general interest in topics and possible motivations for algorithmic policy changes, can help interpret changes in residuals.

The three branches consist of the following features.

- **Caption branch:** While each model is trained on separate groups, which separates topics of captions, previous literature has shown that other variables such as emotions and calls to action, drive user engagement (Chae, 2021; Zhao et al., 2021). Accordingly, I embed the caption of each post into 384 dimensions using SBERT paraphrase-MiniLM-L3-v2 (Reimers & Gurevych, 2019). These are passed to a linear layer that halves the dimension of these embeddings, which are then passed to an LSTM. The LSTM has a memory of the past fourteen posts, so the branch is working on the reduced embeddings of the past fourteen posts.

- **Time branch:** The time branch uses an LSTM architecture with a memory of the past fourteen posts, meaning the LSTM is working on the features described below for the past fourteen posts.

- **Rolling average:** The time series data is heavily auto-regressive, meaning that the number of log likes on previous posts is a good predictor of number of log likes for the current post. Thus, the rolling average of the previous fourteen posts of an account is included as a feature. This value accounts for the popularity of a certain account (e.g. any post from an account with few million followers is likely to get more likes than an account with a few thousand) and trends within an account (e.g. sudden virality often leads to higher engagement with future account content). Since the past fourteen posts are passed to the LSTM, the fourteen previous rolling averages, which were taken over a period of fourteen posts, is passed to the LSTM.
- **Trigonometric time features:** User engagement depends on user schedules, which are built around months (for holidays and seasons), day of week (work days, weekends), and hour (working hours, free time). Existing research from brand engagement tools and academia confirm the interaction effect between day of week and hour ("The Best Time to Post on Social Media in 2025: Times for Every Major Site", 2024, "Best times to post on social media in 2024", 2024, Zhao et al., 2021). Therefore, six of the eight time features are sine and cosine features of the hour of day, day of week, and month of year (e.g. $sin(2 * \pi * hour/24)$).
- **First and Second Derivatives:** To account for changes in rate of growth, I use the first and second derivative of the change in log likes over time. Derivatives capture information about the rate of change and contain information about expectations of future growth. While deep networks should be able to find derivatives, if useful, on their own, my architectures only have a few layers. In testing, models using these derivatives performed meaningfully better than models without.
- **Time delta:** The time series data for each account is highly irregular, since accounts do not post on a schedule consistent with each other and (often) themselves. The two main methods to deal with irregular time series data is imputation of missing values to create standard time steps, or develop a model that handles irregular data. Since my data has a high percent of missing data due to high variation in posting frequency within and across accounts, I focus on flexible methods. According to Weerakody et al., 2021, specific strategies for these models include input augmentation (adding features that indicate missingness or time delta), time decay factors, and ordinary difference equations (ODE) that parameterize the derivative of the hidden state. The time decay method assumes that previous observations from a long time ago are less influential. Since previous post engagement often leads to long term followers, the time decay factor

is weaker for my data. Giving the model time delta allows the model to learn how time delta may affect engagement. For posts with -1 likes, while I do not include their data in the dataset, I do use them to derive time delta features (i.e. time delta is time since last post in collection data, not time since last post in training / testing data).

- **Username branch:** Since accounts within groups behave differently, and may be algorithmically reduced or amplified differently (i.e. accounts with millions versus thousands of followers, accounts that post ten times a day versus twice a month), I add one-hot encoded dummy variables for usernames. These dummy variables are passed to a linear layer, and the outputs are averaged across the fourteen timestamps. Since the usernames are the same for each timestamp, taking the average or last step is equivalent.

The caption branch passes the 384 dimension SBERT embeddings to a linear layer, then a long short-term memory (LSTM) network, where the last item is taken as the output. The time branch passes the ten time features (since/cosine of hour/week/month, rolling average, derivatives, and time delta) to an LSTM and the last item is taken as the output. The username branch takes the 312 dummy variables (for 312 unique usernames) through a fully connected linear layer, and the mean across the last fourteen posts is taken. Since the posts all come from the same account, using mean or last item is interchangeable for the usernames. The three models successively add branches: caption; caption and time; caption, time, and usernames. All models share a final sequential network, where the output of the branch(es) are concatenated if needed, and passed through two hidden layers with ReLU activations. Only groups with at least 1,000 posts in the before dataset are trained. All models use mean squared error loss, the Adam optimizer with 0.001 learning rate, and 20 epochs.

## 2.4 Comparison of Architectures

Out of the three architectures, the two-branch model was chosen for further use and analysis. On the control split, it had the smallest median validation loss across groups. While the residuals for before and after align well for all three model architectures for the control split, a more volatile data split shows that the three branch model overfits for data before, and does not generalize well to data after.

For the model drift test, I compare the residuals of the validation set from the before data to the residuals of the after data so the test is only comparing data the model was not trained on. Figure 9 shows that the distribution of residuals from validation and all before

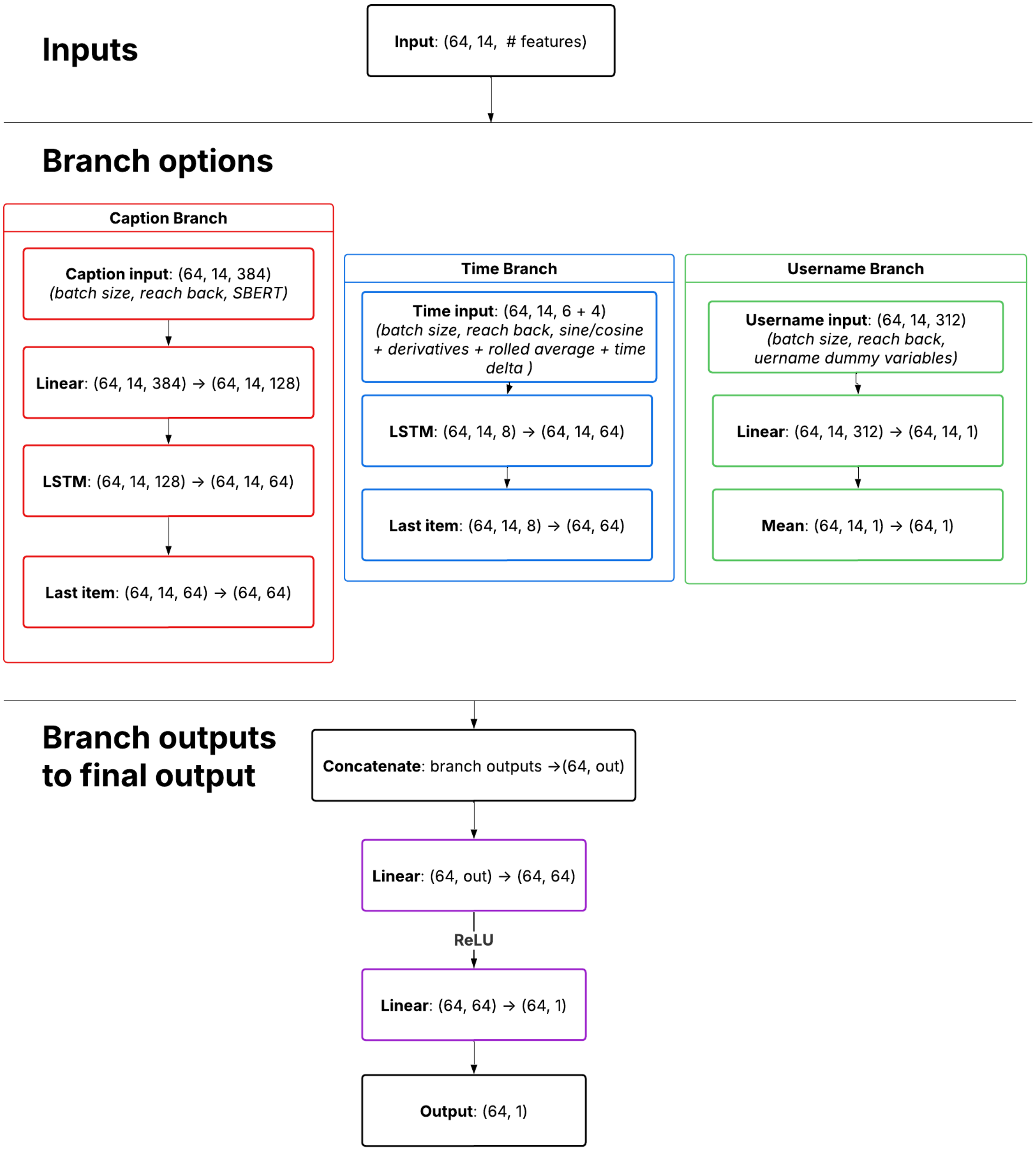
Inputs
Input: (64, 14, # features)
Branch options
Caption Branch
Caption input: (64, 14, 384)
(batch size, reach back, SBERT)
Linear: (64, 14, 384) → (64, 14, 128)
LSTM: (64, 14, 128) → (64, 14, 64)
Last item: (64, 14, 64) → (64, 64)
Time Branch
Time input: (64, 14, 6 + 4)
(batch size, reach back, sine/cosine + derivatives + rolled average + time delta )
LSTM: (64, 14, 8) → (64, 14, 64)
Last item: (64, 14, 8) → (64, 64)
Username Branch
Username input: (64, 14, 312)
(batch size, reach back, uername dummy variables)
Linear: (64, 14, 312) → (64, 14, 1)
Mean: (64, 14, 1) → (64, 1)
Branch outputs to final output
Concatenate: branch outputs →(64, out)
Linear: (64, out) → (64, 64)
ReLU
Linear: (64, 64) → (64, 1)
Output: (64, 1)

Figure 6: Architecture of the three models

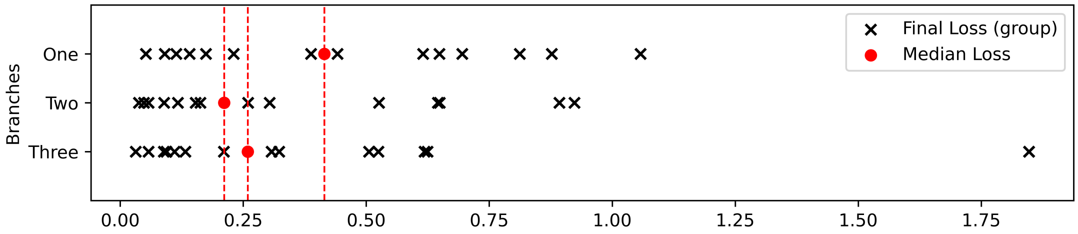

Figure 7: Validation loss
A plot of validation loss by model architecture. Each point represents the validation loss of model trained and tested on one group. The red dot shows each architectures' median loss.

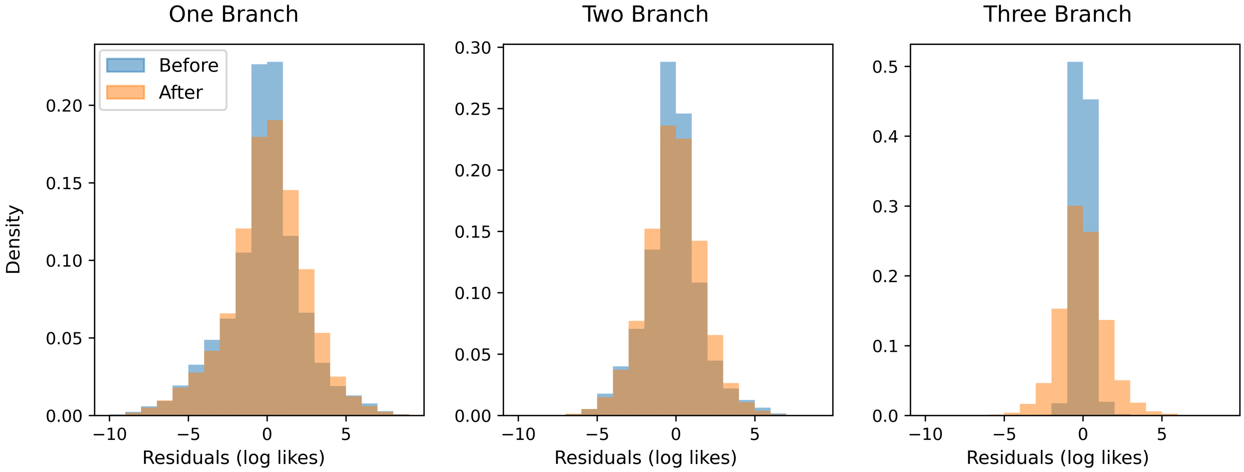

Figure 8: Comparison of residuals across models
The three histograms compare the before and after residual distribution for each of the model architectures.

data are similar. The distribution of residuals for the after data is not as concentrated as data before, which is expected, but it is centered around zero.

# 3 Results

## 3.1 Control

Many of the model drift t-statistics for the groups using the control split were insignificant. Figure 10 shows the relationship between model drift t-statistic and residual distributions. The left bar graph plots model and data drift t-statistics, with transparent values representing insignificant values. The blue bars represent the model drift t-statistics, and the black

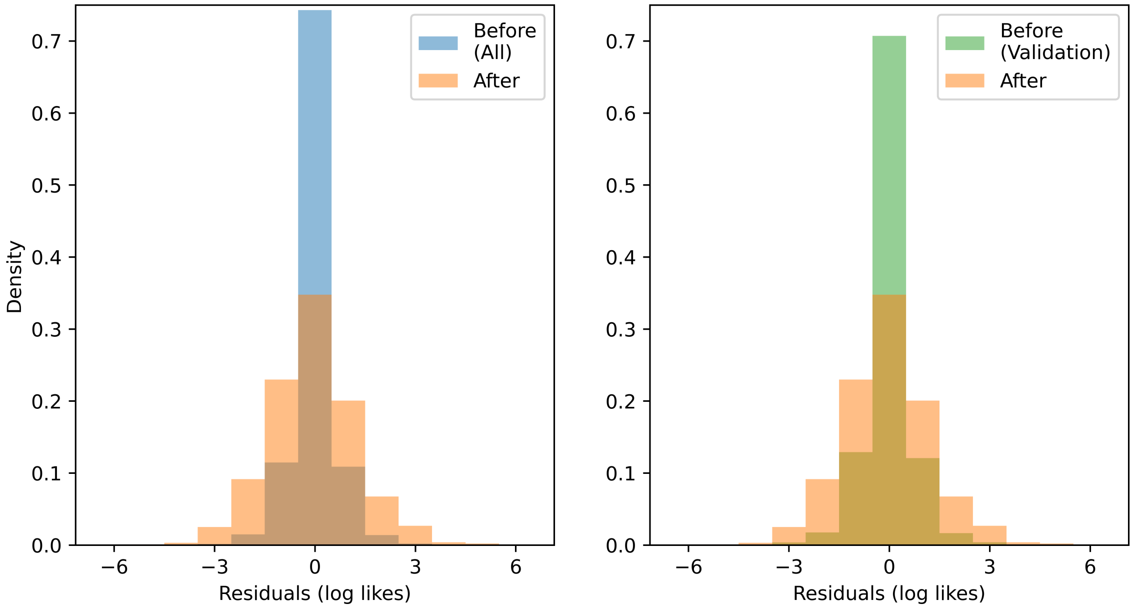

Figure 9: Comparison of residuals across validation and all before data
The two histograms compare the residuals of after data to the residuals of validation/all before (training and validation) data.

line represents data drift t-statistics. Groups with insignificant model drift t-statistics include celebrities, gun, and LGBTQ. The plots on the right show the comparison of residuals from the data before and after. For these groups, the distribution of residuals from the data after are centered around zero. Groups with significant positive model drift t-statistics, such as cooking and right-wing health content show that the residuals after are shifted towards the left, with more negative values. For these groups, the model trained on before data has a tendency to overestimate log likes for the data after. Brands and Republicans have negative model drift t-statistics, indicating the residuals are more positive, and the model underestimates log likes for the data after.

Almost all data drift t-statistics are insignificant. The sign of the model and data drift t-statistics are in agreement for news and Republicans, and disagree for Democrats. News has positive t-statistics, indicating the mean log likes in the after data is less than that of the before data and the model overestimates log likes in the data. By comparing the sign of these two t-statistics, it's likely that the model was overestimating log likes for data after because the data after had less log likes than the data before, and the model was not trained on data after. Similarly, the Republicans have negative t-statistics, indicating the model was underestimating log likes likely because the data after had more log likes than the data before. In contrast, Democrats have a negative data drift t-statistic and a positive model drift t-statistic, meaning that even though the data after had more log likes than the data before, the model still overestimated log likes for data after, which may indicate algorithmic reduction.

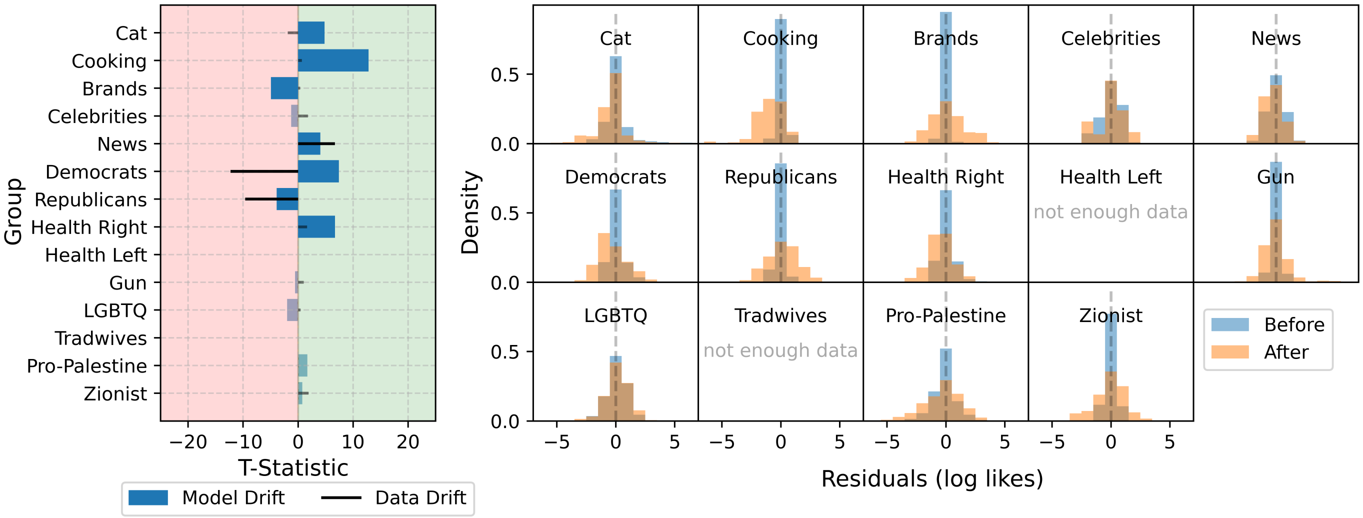

Figure 10: T-Statistics and Histograms on Control Split
The left bar graph plots model and data drift t-statistics. Blue bars represent model drift and black lines represent data drift. Transparent values, bar or line, represent insignificant values. Histograms on the right show the comparison of model residuals from the data before and after.

## 3.2 Split One

The model drift t-statistics for the models trained on -4, -3, -2, and -1 months before the February 2024 policy change are shown in Figure 11. Similar to Figure 10, the black lines show the data drift t-statistics, with the before set corresponding to its respective bar color. Celebrities, news, right-wing health content both have positive model and data drift t-statistics, indicating that these categories have a smaller mean log likes in the data after than the data before, which may explain why the model overestimates log likes for the after set. In contrast, the Democrats and Republicans have negative data drift t-statistics, indicating a higher mean after than before, but positive model drift t-statistics, indicating the model overestimates on data after. This may point to algorithmic reduction, which lines up with the policy change of February 2024 that stated it includes content related to laws and elections. The disagreement in sign is present for all four before dates, indicating that measures to algorithmically reduce the type of content Democrats and Republicans post may have started before October 2023.

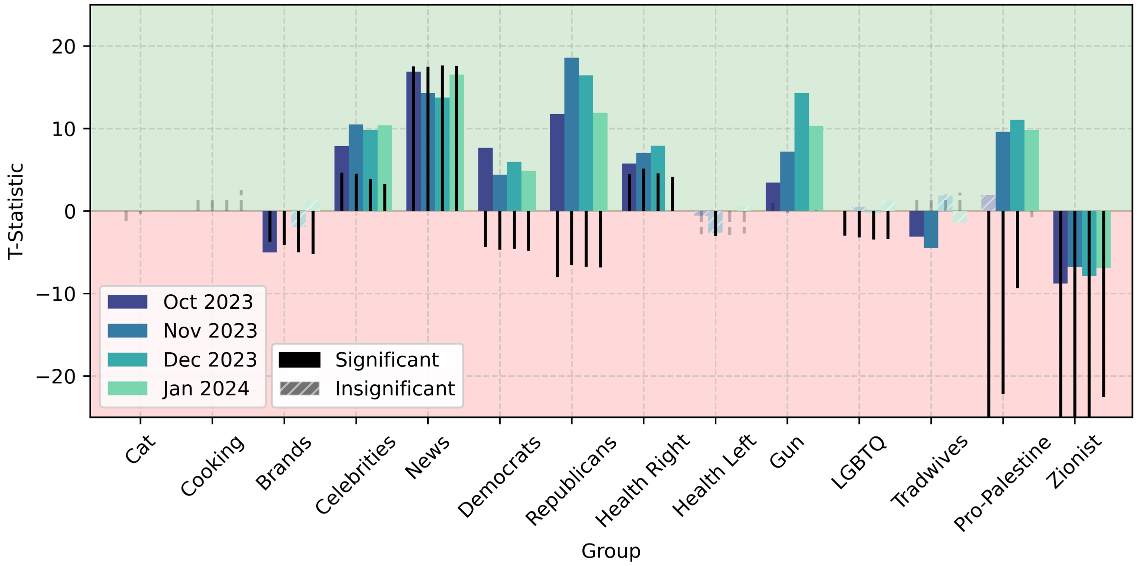

Figure 11: T-Statistics of Data Split One
This graph depicts model and data drift t-statistics comparing data four/three/two/one months before February 2024 to data after February 2024. Bars represent model drift t-statistics, and black lines represent the corresponding data drift t-statistic.

Most interestingly, the model drift t-statistics behave differently for pro-Palestine versus Zionist content. All data drift t-statistics are negative for both groups, which is expected due to the large jump in engagement after October 2023. The model drift t-statistics for Zionist content are also negative. However, the model drift t-statistics for pro-Palestine content is insignificant, then positive as more data is included in the before set. Figure 12 shows a more granular view of these model drift t-statistics, with models trained on data including about 0 weeks, 1 day, 1 week, and 2 weeks after October 7, 2023. The first two model drift t-statistics for pro-Palestine content are insignificant, and the last two are positive, while all model drift t-statistics for Zionist content are negative. The agreement in negative sign for Zionist content is expected: engagement significantly increased after October 2023, and most of the training data happens before the jump in engagement, and so the model continues to predict lower values of log likes. This is true for models trained on data including zero, one, two weeks, and one, two, three months after October 2023. This is supported by Figure 13, which shows model drift t-statistics for before splits that happen much later, and thus include more data after October 2023. Here, the model has seen more data from after the jump in engagement, and the model drift residuals for Zionist content are insignificant, indicating there is not a bias for underestimating log likes for the data after. It's interesting to note that the data drift t-statistics remain negative for Zionist content, but become insignificant for pro-Palestine content. This could be a reflection of the impact algorithmic changes have on user interests. Additionally, frequency of pro-Palestine posts decreased soon after October 2023 (see Figure 4), while Zionist post frequency remained stable after the initial jump, which may also affect user interests.

While it would be expected that the models would exhibit similar behavior for the pro-Palestine content due to a similar jump in interest at the same time, the models actually begin to overestimate log likes for the data after as soon as one week after October 7th 2023, even though the log likes in the after sets have a greater mean than the log likes in the before sets. Thus, the models' overestimation is probably not due to behavioral drift, but rather system drift, indicating signs of algorithmic reduction.

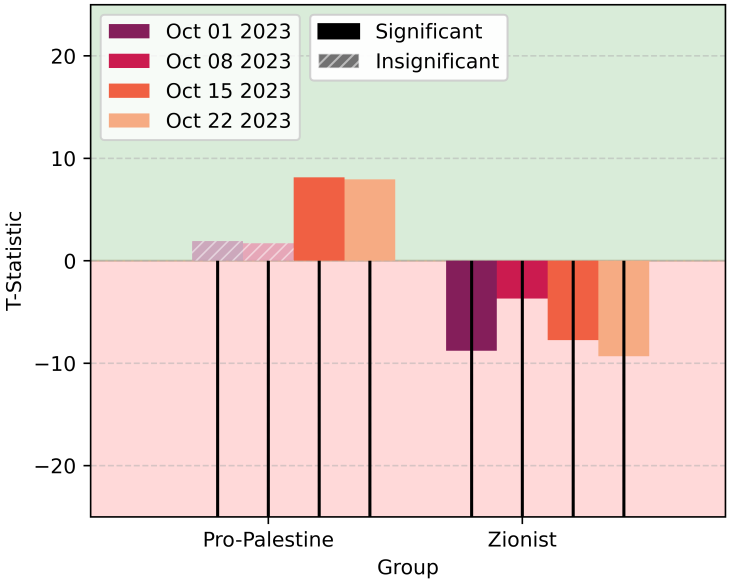

Figure 12: Weekly T-Statistics for October 2023
This graph depicts the model and data drift t-statistics for pro-Palestine and Zionist content comparing data before each week of October 2023 to after February 2024.

## 3.3 Split two

The model drift t-statistics for the models trained on -4, -3, -2, and -1 months before the January 2025 policy change are shown in Figure 13. While this policy change was likely motivated by a single event, Trump's second election, the issues related to the event are not well captured by any topic. However, Instagram's new policy did explicitly change hate speech rules surrounding gender and sexuality. It might be expected that categories related to gender and sexuality may experience system shift, but the model drift t-statistics of

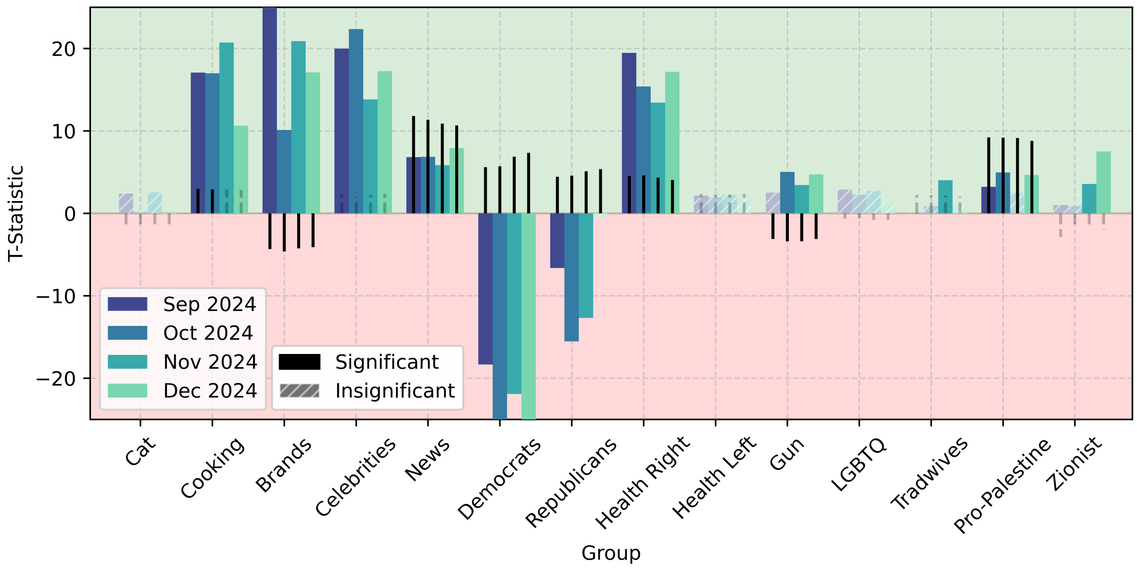

Figure 13: T-Statistics of Data Split Two
This graph depicts model and data drift t-statistics comparing data four/three/two/one months before January 2025 to data after January 2025. Bars represent model drift t-statistics, and black lines represent the corresponding data drift t-statistic.

tradwives and LGBTQ content have mostly insignificant model and data drift t-statistics. Accordingly, the mean log likes in the before and after sets for these categories did not meaningfully change, and models trained on these categories did not have a tendency to under- or over- estimate log likes for the after set compared to the before set. However, t-statistics for gun content disagree, and suggest that the content may have been algorithmically reduced.

The strongest changes seen between Figure 11 and Figure 13 are in Democrat, Republican, pro-Palestine, and brand content. The disagreement in sign for brand content seems to suggest algorithmic reduction. Model drift t-statistics for pro-Palestine content continue to stay positive, but the data drift t-statistics are also positive here. The engagement with pro-Palestine content has dropped, and the model may now be overestimating for that reason. If algorithmic reduction did occur, this could show the successful use of algorithmic reduction to make a topic less relevant. Democrats and Republicans have completely flipped signs, with negative model drift t-statistics and positive data drift t-statistics, suggesting that the content may now be algorithmically amplified. While this policy change stated that such content would no longer be algorithmically reduced, it did not suggest it would be algorithmically amplified. However, such content may have been algorithmically amplified due to the relevance of the US presidential election.

## 4 Discussion and Conclusion

I designed two tests to compare model drift and data drift in order to disentangle and examine changes in Instagram's algorithmic reduction and amplification policies away from changes in user behavior. While the two mutually influence each other, comparing t-statistic sign from the model and data drift tests capture suggestions of algorithmic reduction or amplification. To investigate two policy changes Instagram announced, I had three main data splits: a control split on July 2023, a split for the policy change announced February 2nd, 2024, and a split for the policy change announced January 7th, 2025. The data and model drift tests on the control split suggested that content from Republicans increased popularity, news decreased, and Democrats increased in popularity but may have been algorithmically reduced. The split for the first policy change indicates that celebrities, news, and right-wing health content decreased in popularity, politicians (Democrats and Republicans) increased in popularity but may have been algorithmically reduced, and Zionist and pro-Palestine content increased in popularity but only pro-Palestine content showed signs of being algorithmically reduced. A more granular analysis reveals that signs of algorithmic reduction start as soon as one week after October 7th, 2023. The split for the second policy change indicates content from brand accounts increased in popularity but may have been algorithmically reduced, Democrats and Republicans decreased in popularity but may be algorithmically amplified, and pro-Palestine content decreased in popularity and no longer shows signs of algorithmic reduction.

In the end, while cause and effect of user attention and algorithmic shifts are deeply entangled and cannot be unknotted, I was able to use Instagram post data to show changes in engagement in connection to relevant events. Changes in log likes and model bias after the policy announcement in February 2024 suggest that content from politicians may have been algorithmically reduced, which aligns with the policy's announced intention to reduce content ”potentially related to things like laws, elections, or social topics” (“Update on Political Content on Instagram and Threads — about.instagram.com”, 2024). Pro-Palestine content may have been algorithmically reduced as soon as one week after October 7th, 2023, which aligns with accusations of US censorship regarding the genocide in Gaza (Conley, 2024; de Guzman, 2024; Perkins, 2024; Thakker & Lacy, 2024).

The data shows many changes after the policy announcement in January 2025, but there are none widely backed by relevant news and literature. The data split for this policy did only include about two and a half months of data in the ‘after’ set, which may make these results more volatile for two reasons. Since post engagement was collected at a single point in time, likes for posts closer to the collection date may not have stabilized and reached the full amount of likes they will eventually reach. With two and a half months of data, a larger proportion of this data may not have gathered the full amount of likes. Additionally, the shorter time period in the data after may be more sensitive to particular variations connected with time. As a result, although there are enough posts to have significant t-statistics, results may be more volatile due to the shorter time period.

Throughout their existence, social media platforms have worked to position in order to enjoy maximum monetary benefits with minimum social and public responsibility. The current turn to algorithmic moderation systems is the latest method of doing so. Using algorithmic reduction rather than content removal allows platforms to use the black box nature of algorithms to artificially create opacity surrounding reduction (Gillespie, 2022), evade responsibility through "black box gaslighting," (Cotter, 2023), and take advantage of a Foucauldian relationship between platform and creator (Duffy & Meisner, 2023). In relation to my work, the manufactured opacity is important for reducing political and monetary risk. By creating opacity around what content is affected by algorithmic policy changes, Instagram is able to make algorithmic policy changes according to executive interests while deflecting accusations of bias and censorship. Specifically, for the ongoing genocide in Gaza, Instagram denied accusations of bias against pro-Palestine content, stating that their policies are equally applied and "there is no truth to the suggestion that we are deliberately suppressing voice" and pointing to bugs that resulted in reduced reach ("Meta's Ongoing Efforts Regarding the Israel-Hamas War", 2023). With the results of this project, however, there is evidence to suggest that the February 2024 policy was indeed applied unequally with regard to pro-Palestine and Zionist content.

While platforms intentionally manufacture opaqueness surrounding algorithmic reduction and amplification for their benefit, transparency is crucial for holding platforms responsible to the large scale political and social effects they have on the world. There is an extended history of the ways the algorithmic recommendation systems on social media has shaped political reality in many spheres: Facebook contributed to doubt about severity of COVID and safety of vaccines ("The Facebook Files", 2021), the global support garnered for the Arab Spring on multiple platforms (Smidi & Shahin, 2017), echo chambers used as tools of propaganda that precipitated in the genocide of the Rohingya people ("Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations – new report", 2022). The epistemic power imbalance between platform and user allows platforms to evade responsibility for contributions of real world events. While it remains true that algorithms, including algorithmic recommendation systems currently remain black boxes, their outputs and effects of reduction and amplification are not. Large scale efforts to understand censorship in the form of algorithmic reduction remains extremely sparse. My project serves, in small part, to fill this gap and chip away at the epistemic power imbalance.

Future research could focus on more robust methods of detecting algorithmic reduction and amplification by collecting more data and focus on methods to disentangle data drift and system drift. Although I use a large number of posts, the methods require multiple divisions of these data points, from individual posts to accounts to groups to before/after and train/validate sets, reducing the number of posts in each bucket. While user engagement and algorithmic changes are inherently tied to each other, outside events, such as motivating events and policy changes, do act as natural experiments, and one comes before the other. Perhaps a major world event happens, and engagement changes, then algorithmic suggestions change. Perhaps a policy change happens, and algorithmic suggestions change, then long term engagement changes. There is a gap between the two, and more data could allow for better operationalization to capture the effect of a natural experiment, such as comparing data after the last policy change to before the next one to data right after the policy change in question. Since my method includes all data before or after a certain date, other relevant events could be mixed into the changes seen. Accordingly, more granular time scales for model drift and data drift tests could reveal more precise changes, as shown in Figure 12.

Additionally, my data collection limits the study in two forms. While removal of content is not the object of study, removal of content does happen, and means the data is subject to selection bias, since I can only collect posts that have not been removed. Recent leaked data confirms that Meta responds to government requests for content removal. Since October 7, 2023, the company significantly expanded automated content removal and has complied with 94% of removal requests issued by Israel, the biggest originator of removal requests globally (Ahmed et al., 2025). The method of collection also limits the accuracy of algorithmic reduction detection. It is expected that a post gets most of its likes when fresh, and decreases over time, eventually to a stopping point. Many creators notice they have been algorithmically reduced when they do not receive the expected interaction (views, likes, comments) during this fresh period. Since I only visited each post once, this data cannot be used to look for algorithmic reduction by time series data of individual posts.

As worries about the right to free speech and protest grow in the US, transparency around how content is amplified and reduced on platforms become even more critical. The results of my research suggest that Instagram, a major social media platform in the US, has shown signs of algorithmic reduction of content that does not align with US government interests. Since transparency goes against platform motivations, research to bring transparency and level the epistemic power is greatly needed.

## Data and Code Availability Statement

Researchers interested in access to the data may contact Grace Shao at graceshao.200@proton.me. Code for analysis is provided in this GitHub repository.

## References

Ahmed, W., Rodelo, N., Grim, R., & Hussain, M. (2025, April). *Leaked data reveals massive israeli campaign to remove pro-palestine posts on facebook and instagram* [[Accessed 24-04-2025]].

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211–236.

Ames, M. G. (2018). Deconstructing the algorithmic sublime.

The Best Time to Post on social media in 2025: Times for every major site. (2024). *Buffer*. https://buffer.com/resources/best-time-to-post-social-media/

Best times to post on social media in 2024. (2024). *Sprout Social*. https://sproutsocial.com/insights/best-times-to-post-on-social-media/

Broniatowski, D. A., Jamison, A. M., Johnson, N. F., Velasquez, N., Leahy, R., Restrepo, N. J., Dredze, M., & Quinn, S. C. (2020). Facebook pages, the "disneyland" measles outbreak, and promotion of vaccine refusal as a civil right, 2009–2019. *American journal of public health*, *110*(S3), S312–S318.

Chae, M.-J. (2021). Driving consumer engagement through diverse calls to action in corporate social responsibility messages on social media. *Sustainability*, *13*(7), 3812.

China to implement new regulation on algorithm recommendation services [[Accessed 24-04-2025]]. (2022). *www.gov.cn*. https://english.www.gov.cn/news/topnews/202201/04/content_WS61d3f8fbc6d09c94e48a31d1.html.

China to tighten regulation of algorithms related to internet information services [[Accessed 24-04-2025]]. (2021). *www.gov.cn*. https://english.www.gov.cn/statecouncil/ministries/202109/29/content_WS61546088c6d0df57f98e10f2.html.

Cole, S. (2018, July). *Where Did the Concept of 'Shadow Banning' Come From? — vice.com* [[Accessed 27-04-2024]].

Conley, J. (2024). Romney admits push to ban tiktok is aimed at censoring news out of gaza [[Accessed 22-03-2025]]. *Common Dreams*. https://www.commondreams.org/news/mitt-romney-tiktok.

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. https://doi.org/10.1080/1369118X.2021.1994624

de Guzman, C. (2024). Instagram's Political Content Limit: Everything to Know [[Accessed 22-04-2024]]. *Time*. %5Curl%7Bhttps://time.com/6960587/meta-instagram-political-content-limit-off-setting-default/%7D

Diamond, L., & Plattner, M. F. (2012). *Liberation technology: Social media and the struggle for democracy*. JHU Press.

Duffy, B. E., & Meisner, C. (2023). Platform governance at the margins: Social media creators' experiences with algorithmic (in)visibility. *Media, Culture & Society*, *45*(2), 285–304. https://doi.org/10.1177/01634437221111923

Gillespie, T. (2018, January). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. https://doi.org/10.12987/9780300235029

Gillespie, T. (2022). *Yale Journal of Law and Technology*, *24*, 476.

Horten, M. (2024). Algorithms patrolling content: Where's the harm? *International Review of Law, Computers & Technology*, *38*(1), 43–65.

Horwitz, J., Hagey, K., & Glazer, E. (2023). Facebook wanted out of politics. it was messier than anyone expected. *Wall Street Journal*. https://www.wsj.com/articles/facebook-politics-controls-zuckerberg-meta-11672929976

Human Rights Due Diligence of Meta's Impacts in Israel and Palestine in May 2021 [[Accessed 05-01-2025]]. (2022). *Business for Social Responsibility*. https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf.

Israel/Palestine: Facebook Censors Discussion of Rights Issues [[Accessed 05-01-2025]]. (2021). *Human Rights Watch*. https://www.hrw.org/news/2021/10/08/israel/palestine-facebook-censors-discussion-rights-issues.

Ivanski, C., Lo, R. F., & Mar, R. A. (2021). Pets and politics: Do liberals and conservatives differ in their preferences for cats versus dogs? *Collabra: Psychology*, *7*(1), 28391.

Kaplan, J. (2025). More Speech and Fewer Mistakes [[Accessed 07-03-2025]]. *Meta*. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes.

Katzenbach, C. (2021). "AI will fix this" – The Technical, Discursive, and Political Turn to AI in Governing Communication. *Big Data & Society*, *8*(2), 20539517211046182. https://doi.org/10.1177/20539517211046182

King, G., Pan, J., & Roberts, M. E. (2013). How censorship in china allows government criticism but silences collective expression. *American Political Science Review*, *107*(2), 326–343. https://doi.org/10.1017/S0003055413000014

King, G., Pan, J., & Roberts, M. E. (2014). Reverse-engineering censorship in china: Randomized experimentation and participant observation. *Science*, *345*(6199), 1251722.

Lavietes, M. (2025). Meta's new hate speech guidelines permit users to say lgbtq people are mentally ill. *National Broadcasting Company*. https://www.nbcnews.com/tech/social-media/meta-new-hate-speech-rules-allow-users-call-lgbtq-people-mentally-ill-rcna186700

Le Merrer, E., Morgan, B., & Trédan, G. (2021). Setting the record straighter on shadow banning. *IEEE INFOCOM 2021-IEEE conference on computer communications*, 1–10.

Maiberg, E., & Koebler, J. (2025). This 'College Protester' Isn't Real. It's an AI-Powered Undercover Bot for Cops [[Accessed 24-04-2025]]. *404 Media*. https://www.404media.co/this-college-protester-isnt-real-its-an-ai-powered-undercover-bot-for-cops/.

Meta's Ongoing Efforts Regarding the Israel-Hamas War [[Accessed 05-01-2025]]. (2023). *Meta*. https://about.fb.com/news/2023/10/metas-efforts-regarding-israel-hamas-war/.

Morozov, E. (2012). *The net delusion: The dark side of internet freedom*. PublicAffairs.

*Myanmar: Facebook's systems promoted violence against rohingya; meta owes reparations – new report* [[Accessed 24-04-2025]]. (2022, September).

Nicholas, G. (2022). Shedding light on shadowbanning. *Center for Democracy & Technology. https://cdt. org/insights/shedding-light-on-shadowbanning*.

Perkins, T. (2024). Anti-defamation league ramps up lobbying to promote controversial definition of antisemitism [[Accessed 24-04-2025]]. *The Guardian*. https://www.theguardian.com/us-news/article/2024/may/15/adl-lobby-antisemitism-definition.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. http://arxiv.org/abs/1908.10084

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Smidi, A., & Shahin, S. (2017). Social media and social mobilisation in the middle east: A survey of research on the arab spring. *India Quarterly*, *73*(2), 196–209.

Stpanov, A., & Gupta, A. (2021). Reducing Political Content in News Feed — Meta — about.fb.com [[Accessed 22-04-2024]]. *Meta*. https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed.

Telford, T. (2021). Facebook moves to scale down political content. *Washington Post*. https://www.washingtonpost.com/business/2021/02/10/facebook-political-content/

Thakker, P., & Lacy, A. (2024). In no labels call, josh gottheimer, mike lawler, and university trustees agree: Fbi should investigate campus protests [[Accessed 24-04-2025]]. *The Intercept*. https://theintercept.com/2024/05/04/josh-gottheimer-mike-lawler-campus-protests/.

The Facebook Files. (2021). *Wall Street Journal*. https://www.wsj.com/articles/the-facebook-files-11631713039

Update on Political Content on Instagram and Threads — about.instagram.com [[Accessed 22-04-2024]]. (2024). *Meta*. https://about.instagram.com/blog/announcements/continuing-our-approach-to-political-content-on-instagram-and-threads.

Weedston, L. (2025). ‘i’m filled with the most impotent rage’: Americans are becoming radicalized after seeing grocery, healthcare costs in china [[Accessed 24-04-2025]]. *daily dot*. https://www.dailydot.com/news/americans-radicalized-by-rednote/.

Weerakody, P. B., Wong, K. W., Wang, G., & Ela, W. (2021). A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing*, *441*, 161–178.

World court finds israel responsible for apartheid [[Accessed 30-03-2025]]. (2024). *Human Rights Watch*. https://www.hrw.org/news/2024/07/19/world-court-finds-israel-responsible-apartheid

Zeng, J., & 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. https://doi.org/10.1002/poi3.287

Zhao, Z., Liu, Y., Wang, J., Wang, B., & Guo, Y. (2021). Association rules analysis between brand post characteristics and consumer engagement on social media. *Engineering Economics*, *32*(4), 387–403.

Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. *Annual review of economics*, *12*(1), 415–438.

# Appendix

## A: Usernames

### A:.1 Successfully Collected

**Cat:** nala_cat, realgrumpycat, smoothiethecat, catconworldwide, maple.cat, catinberlin, catladybox, triumphant_teagan, aliencatmatilda, mollymollzthetabby, my_lulu_cat_, my_furry_babies, monicasisson, allcreaturestv, moana.and.snapple, thatcatconrad, catsvscancer, baili_the_cat, panther.cat, butter_the_siberian, trippy.tails, bearbear.cat, mochicat168, bellina_kitty_cat, rainerogers, ambrosepets, realbadgalrhirhi, pawaiihub, denbo_nish, siberian.milo
**Cooking:** halfbakedharvest, ketosnackz, nourishing, olivia.adriance, chloecleroux, geoffreyzakarian, winnyhayes, alyssacoadynutrition, liveeatlearn, thefoodnanny, femalefoodie, sherryhour, allinspiredwellness, mytoddlerskitchen, reciperunner, lydialove98, mleroehler, anniesfinds_, littlespoonfarmblog, perfectsupplements, feelgoodwithfi, yourstrulyani, thismomentinthyme, leahmariestack, through.manals.lens, recipesfrommichelle, mccauley_tawpash, mintandclove, ashleighbovard, neeleman_food_
**Brands:** starbucks, apple, amazon, google, microsoft, louisvuitton, prada, target, mcdonalds, kfc, dominos, spotify, instagram, xbox, disney, pixar, hulu, nikefootball, adidasfootball, underarmour, victoriassecret, calvinklein, tommyhilfiger, cartier, gopro, canon, nikon, sonymusic, warnerrecords, cosmopolitan
**Celebrities:** cristiano, leomessi, selenagomez, therock, kyliejenner, arianagrande, kimkardashian, beyonce, khloekardashian, justinbieber, kendalljenner, taylorswift, jlo, nickiminaj, kourtneykardash, mileycyrus, katyperry, zendaya, kevinhart4real, kingjames, ddlovato, badgalriri, champagnepapi, ellendegeneres, k.mbappe, billieeilish, lalalalisa_m, vindiesel, shraddhakapoor, priyankachopra, narendramodi, shakira, davidbeckham, jennierubyjane, aliaabhatt
**Democrats:** joebiden, kamalaharris, barackobama, michelleobama, berniesanders, aoc, chuckSchumer, elizabethwarren, amyklobuchar, corybooker, chrismurphyct, repjerrynadler, repkatieporter, repvaldemings, repdebhaaland, repmarkpocan, repjimmygomez, repdavidcicilline, repdonbeyer, repderekkilmer, repsusielee, repkathleenrice, repjimhimes, repgregstanton, repjuanvargas, repmikethompson, repjimlangevin, repjimcosta, repjimmygomez, repjoshharder, repkatiehill
**Republicans:** mikepence, realdonaldtrump, lindseygrahamsc, tedcruz, mittromney, pauldavisryan, ronjohnsonwi, marcorubio, newtgingrich, speakermccarthy, johncornyn, rondesantis, kristinoem, stevescalise, markmeadows, elisestefanik, leezeldin, cathymcmorris, replizcheney, repmattgaetz, repbrianmast, repgregpence, repmarkgreen, repbuddycarter, repjeffduncan, repdavidkustoff, reppatfallon, repchrisstewart, repkenbuck

**News:** cspan, bloomberg, businessinsider, vice, reuters
**Gun:** garand_thumb, gunpolicy, gunownersofamerica, sb.tactical, lawtactical, midwestindustries, centuryarms, gundrummer, griffin_armament, down_range_photography, maximdefense, zaffiri.precision, pewpewtactical, killerinnovations, rarebreedfirearms, battleborn, xtechtactical, gregskazphotography, kci_usa, firearmchronicles, patriot_defense_gear, 704_tactical, armedscholaryt, shootersgrill, elevatedsilence, sdgunowners, womenforgunrights, gunownersca, 2arally, rmgo_official
**Health Right:** thetruthaboutcancerttac, elaineshtein, drmercola, joshsfarmersmarket, dr.goodyear, raw_farm_usa, farmmatch, momsacrossamerica, freedom.hill.farm, ilanamuhlsteinrd, jessalyn.randle, bobolinkdairyandbakehouse
**Health Left:** peoplescdc, youlookokaytome, topheravila, longcovidjustice, thaibrows, la.spoonie.collective, transgressivemedicine, itsjiyounkim
**LGBTQ:** them*, themilesmckenna, queer_lective, dylanmulvaney, themme_fatale, chellaman, tanyacompas, mattxiv, sadegiliberti, jessicaoutofthecloset, jake_graf5, raindovemodel, plussizetransguy, trans.ginger, trevorproject, chandlernwilson, blacktranstravelfund, mpjinstitute, gabesdunn
**Tradwives:** ourquaintandcozy, call_mejewels, ballerinafarm, esteecwilliams, classicallyabby, naraaziza, lifewithmrsp, _cynthialoewenseguin
**Pro-Palestine:** jd.moha, mikopeled, wizard_bisan1, jewishvoiceforpeace, plestia.alaqad, belalkh, zein_rahma, eid_yara, nooh.xp, savesilwan, amirgharabawi, saher_alghorra, haneen.maher.salem, palmuseum, yplusmedia, right2edu, uospalsoc, bigbigbigthings, jaxpsn, alaa_fayez.12, bayanpalestine, mohammadhureini, hind.touissate, taniasafi, apc_uk_london, shirien.creates
**Zionists:** israelcc, adielofisrael, _danielbraun, montanatucker, proudzionista, antisemitism, aipac, dahliakurtz, cameraoncampus, mactaskforce, elizabethyounger, lanianpo, henmazzig, freejamshidsharmahd, worldjewishcongress, michahdoot, leetrink, themodernmaccabee, ajewishresistance, zicksworld, abbasez

*removed for having a significantly higher post frequency than the rest of the accounts in the group

### A:.2 Failed Usernames

**Cat:** rajathebengal
**Cooking:** smittenkitchen, 177milkstreet, simplicityandastarter, jamleenbears, carnivore_connoisseur_
**Brands:** dior, hugo, nba, nintendo, rayban
**Celebrities:** snoopdogg, dualipa
**Democrats:** n/a

**Republicans:** potus, tx

**News:** wsj, financialtimes, msnbc, theintercept, aljazeera, foxnews, guardian, time, nytimes, newsweek, bbcnews, nypost, cnn, forbes,politico, huffpost, abcnews, nbcnews, axios, thehill, cbsnews, usatoday, washingtonpost, latimes, thedailybeast, npr, apnews

**Gun:** n/a

**Health Right:** crunchykass, theregenaissance

**Health Left:** clean.air.club, thesicktimes, berlin_buyers_club, thecovidcollection, long_hauler_haven, maskednh, jaydocovid, maskblocseattle, cleartheair.atx

**LGBTQ:** genderlib

**Tradwives:** hannahlee.yoder, simplyalliehomestead, zimcolorado

**Pro-Palestine:** lama_jamous9, jenanmatari, dr.ghassan.as, palestinianyouthmovement, sjp.uo, wizard_bisan2, palestinehouseoffreedom, nadiforpalestine, queersinpalestine, operationolivebranch

**Zionists:** betarworldwide, standwithus, jewishwomen4allwomen, ajc.global, strength4israel, j.majburd, israel365action, bringhomenow, bring.amiram.home.now, kidnappedfromisrael, natashahausdorff, allhostages, bringbackourhearts, thepersianjewess