# Simulating User Watch-Time to Investigate Bias in YouTube Shorts Recommendations

Nitin Agarwal[1,2], Selimhan Dagtas[1], and Mert Cakmak[1]

[1] COSMOS Research Center, University of Arkansas–Little Rock, USA
[2] International Computer Science Institute, University of California, Berkeley, USA

{nxagarwal,sedagtas,mccakmak}@ualr.edu

**Abstract.** Short-form video platforms such as YouTube Shorts increasingly shape how information is consumed, yet the effects of engagement-driven algorithms on content exposure remain poorly understood. This study investigates how different viewing behaviors, including fast scrolling or skipping, influence the relevance and topical continuity of recommended videos. Using a dataset of over 404,000 videos, we simulate viewer interactions across both broader geopolitical themes and more narrowly focused conflicts, including topics related to Russia, China, the Russia–Ukraine War, and the South China Sea dispute. We assess how relevance shifts across recommendation chains under varying watch-time conditions, using GPT-4o to evaluate semantic alignment between videos. Our analysis reveals patterns of amplification, drift, and topic generalization, with significant implications for content diversity and platform accountability. By bridging perspectives from computer science, media studies, and political communication, this work contributes a multidisciplinary understanding of how engagement cues influence algorithmic pathways in short-form content ecosystems.

**Keywords:** YouTube Shorts, Algorithmic Bias, Recommender Systems, Generative AI, Content Relevance, Watch-Time

## 1 Introduction

Understanding how recommendation systems shape content exposure is vital across fields like computer science, communication, and media studies. Platforms like YouTube Shorts—with rapid, passive engagement and wide reach—offer a key opportunity to study these effects. Examining how short-form algorithms respond to user behavior is crucial for promoting fairness, transparency, and informed media governance.

The rise of short-form video platforms has transformed digital content consumption, with YouTube Shorts emerging as a major force since its launch in 2021. The platform now attracts over 1.5 billion monthly users [7], and its recommendation system plays a central role in shaping what viewers see. While substantial research has explored algorithmic behavior in long-form platforms, there is limited understanding of how these dynamics operate in short-form environments. Recent work has begun to examine bias in Shorts recommendations, highlighting concerns about content drift and visual manipulation [4]. Unlike long videos where engagement involves a range of behaviors, Shorts rely heavily on lightweight interactions such as watch duration and swiping. These platforms demand new approaches to evaluating algorithmic influence, particularly as they continue to expand in reach and cultural relevance.

This study investigates the influence of watch-time behavior on the topical relevance of content recommended by YouTube Shorts. We ask the following research questions:

- **RQ1:** How does user watch-time behavior influence the topical relevance of videos recommended by the YouTube Shorts algorithm over time?
- **RQ2:** Does the use of interest-based watch times amplify recommendation relevance compared to uniform short watch durations?
- **RQ3:** To what extent does the YouTube Shorts algorithm retain topical specificity in recommendations for broader versus more specific topics?

To explore these questions, we conducted an empirical analysis using four thematic datasets: *Broader Russia*, *Russia–Ukraine War*, *Broader China*, and *South China Sea Dispute*. These topics were selected for their geopolitical importance and visibility within online discourse. The Russia–Ukraine conflict represents a major war with widespread global implications [9], while the South China Sea dispute involves contested maritime claims affecting regional security and international trade [2]. These cases provide meaningful testbeds for understanding how recommendation systems handle both general and specific narratives within politically sensitive domains.

## 2 Literature Review

Understanding the dynamics of YouTube's recommendation system, particularly within short-form formats like YouTube Shorts, requires a multidisciplinary perspective that spans computer science, communication studies, and social computing. Prior work has established that watch-time is a key metric driving YouTube's algorithm, gradually replacing earlier click-based models. Covington et al. [5] describe YouTube's two-stage recommendation pipeline, emphasizing its shift toward engagement-optimized learning to maximize user retention. However, this engagement-centric approach raises concerns about personalization loops and algorithmic bias. Studies have shown that recommender systems can reinforce user preferences, contributing to echo chambers and the narrowing of informational diversity [1].

These issues are particularly pressing in politically sensitive or contested information spaces. Understanding how algorithms react to different forms of engagement is critical for assessing their role in shaping public discourse. Network-based methodologies provide tools to examine how content structures evolve in these systems. Community detection techniques such as the Louvain algorithm

[3] have been used to reveal clustering and content silos in recommendation networks, illuminating how algorithmic curation can both reflect and reinforce ideological boundaries. Such structural insights are complemented by recent work in social media studies that highlights how recommender systems influence user exposure and content propagation [10].

While prior work has analyzed recommendation systems and content networks, few studies have simulated how variations in watch-time affect algorithmic behavior. This study fills that gap by modeling fast skipping versus longer viewing within YouTube Shorts, revealing how watch-time alone can steer topical relevance and content trajectories. Our approach combines simulation with network-aware analysis, offering insight into how attention patterns shape personalization and media exposure.

## 3 Methodology

In this section, we present the simulation framework used to examine how varying watch-time behaviors influence the relevance of YouTube Shorts recommendations.

### 3.1 Data Collection

To examine how YouTube Shorts responds to different watch-time behaviors, we collected recommendation data across four themes: Russia–Ukraine War, Broader Russia, South China Sea Conflict, and Broader China. This design contrasts specific, fast-evolving geopolitical conflicts with broader, more stable narratives. Keywords were curated from news sources, policy reports, and academic literature, then refined using language models. For example, Russia–Ukraine terms focused on military and humanitarian developments [14], while South China Sea keywords emphasized maritime disputes [12]. Broader themes reflected topics such as Soviet nostalgia [8] and China's global posture [6]. These keywords guided both search and analysis, allowing systematic comparison of algorithmic responses across topic types. Representative keywords are shown in Table 1.

**YouTube Shorts and Recommendation Collection** — Since the YouTube Data API v3 does not support Shorts, we used APIFY's YouTube Scraper [13] to collect 1,000 Shorts seed videos per topic, totaling 4,000 videos. To capture recommendation dynamics, we developed a custom Selenium-based scraper, as no existing tool collects Shorts recommendations directly. Each seed was tested under two viewing conditions: (1) minimal viewing (3 seconds) and (2) interest-based viewing (3, 15, or 60 seconds), with durations determined by title relevance using a generative AI method (see Section 3.2). Browsing sessions were isolated and extended to a depth of 50 to simulate user scrolling behavior. In total, the framework collected 400,000 recommended videos, yielding 404,000 Shorts across all conditions and topics.

**Table 1.** Summary of Keywords Used for Data Collection. Representative keywords are shown for each dataset.

| **Dataset** | **Keywords (Selected)** |
|---|---|
| Russia-Ukraine | Ukraine frontline update, Russia missile attack, NATO and Ukraine, Ukraine war explained, Ukraine refugee crisis, Russian war crimes Ukraine, Crimea missile strike, Ukraine reconstruction plan. |
| Broader Russia | Russia global influence, Russia-China alliance, Russia BRICS expansion, Russkiy Mir ideology, Russia propaganda abroad, Soviet Union expansion, Russia soft power diplomacy, Russia foreign policy strategy. |
| South China Sea | South China Sea dispute, China nine-dash line, Freedom of navigation operations, Scarborough Shoal standoff, China maritime militia, ASEAN South China Sea talks, China island militarization, US Navy South China Sea. |
| Broader China | China global expansion, Belt and Road Initiative, China military buildup, China debt trap diplomacy, Confucius Institutes controversy, China cyber warfare capabilities, China United Nations influence, China global leadership ambitions. |

### 3.2 Relevancy Measurement

To estimate the topical relevance of recommended YouTube Shorts, we employed generative AI models to score video titles during the scraping process. This automated scoring guided simulated watch behavior in the recommendation collection. Each model received the following prompt: *You are an expert assistant specializing in assessing how relevant a YouTube video title is to topics surrounding [**X**]. For a given YouTube video title, assign a relevance score: 2 for highly relevant, 1 for somewhat relevant, and 0 for irrelevant.* The placeholder [**X**] was dynamically filled with a short description of the dataset theme (e.g., South China Sea maritime conflict, Russia–Ukraine war, China's global expansion, or Russia's foreign policy strategy), as defined earlier.

**Model Selection and Validation** — To evaluate model performance for this task, we tested LLaMA 3, GPT-4o, and Gemini 1.5 on two public QA/relevance datasets: MS MARCO [11] and WikiQA [15]. As shown in Table 2, GPT-4o achieved the highest average accuracy and was used in our analysis.

**Table 2.** Validation accuracy (%) on publicly available QA/relevance datasets used to benchmark the scoring behavior of large language models.

| **Dataset** | **LLaMA 3** | **GPT-4o** | **Gemini 1.5** |
|---|---|---|---|
| MS MARCO [11] | 71.80 | 80.46 | 69.84 |
| WikiQA [15] | 77.50 | 77.00 | 80.60 |
| **Average** | 74.65 | **78.73** | 75.22 |

## 4 Results

Mean relevance scores were compared across 50 depths in the recommendation chain for four thematic datasets: *Broader Russia*, *Russia–Ukraine War*, *Broader China*, and *South China Sea Dispute*. In both Figure 1 (Russia topics) and Figure 2 (China topics), the interest-based watch-time condition consistently produced higher relevance scores across depths compared to the 3-second baseline. These differences emerged early in the chain and persisted through to depth 50.

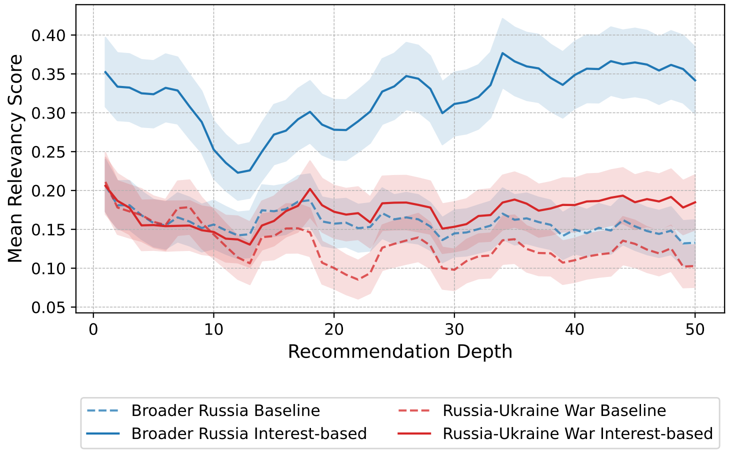

**Fig. 1.** Mean recommendation relevance across depths for the Russia-related topics. The graph compares baseline (3-second skip) and interest-based watch-time conditions for both broader (*Broader Russia*) and more specific (*Russia–Ukraine War*) themes. Shaded regions indicate 95% confidence intervals.

To quantify the overall recommendation quality, we calculated the area under the curve (AUC) for each recommendation path and performed paired t-tests to evaluate differences between conditions. The AUC represents the cumulative relevance that a user encounters across the entire recommendation trail. Results are summarized in Table 3. For example, Broader Russia's AUC increased from 8.10 in the baseline condition to 15.92 in the interest-based condition ($t = 18.404$, $p = 2.42 \times 10^{-46}$). Broader China showed a similarly large shift (5.14 to 12.88, $t = 14.387$, $p = 8.79 \times 10^{-30}$). Smaller but statistically significant gains were observed for the Russia–Ukraine War ($\Delta$AUC $= 1.63$, $p = 1.30 \times 10^{-3}$) and the South China Sea Dispute ($\Delta$AUC $= 0.69$, $p = 2.79 \times 10^{-2}$).

While some increase in relevance is expected when a user watches a video for longer, our findings suggest that even small differences in watch time, ranging from 3 to 60 seconds, produce large and persistent shifts in recommendation paths. On a platform like YouTube Shorts, where all videos are under a minute, these differences are minimal in absolute terms but are treated by the system

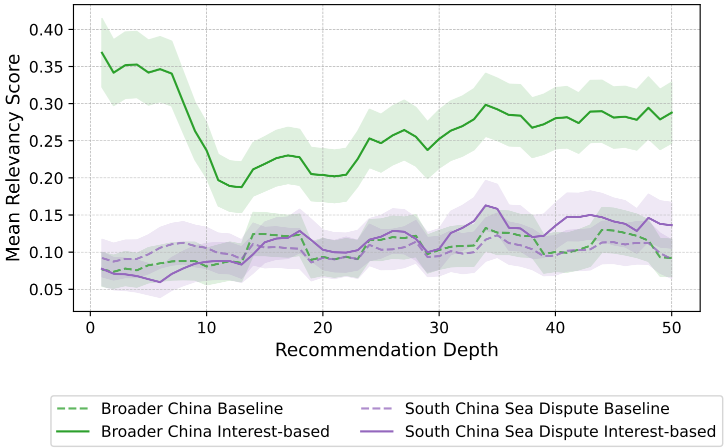

**Fig. 2.** Mean recommendation relevance across depths for the China-related topics. The graph compares baseline (3-second skip) and interest-based watch-time conditions for both broader (*Broader China*) and more specific (*South China Sea Dispute*) themes. Shaded regions indicate 95% confidence intervals.

**Table 3.** AUC comparison between baseline and interest-based watch-time conditions. Paired t-tests compare relevance area under the curve (AUC) across all recommendation depths for each topic.

| **Topic** | **Baseline AUC** | **Interest AUC** | **$\Delta$ AUC** | **p-value** |
|---|---|---|---|---|
| Broader Russia | 8.10 | 15.92 | 7.82 | $2.42 \times 10^{-46}$ |
| Russia–Ukraine | 6.82 | 8.45 | 1.63 | $1.30 \times 10^{-3}$ |
| Broader China | 5.14 | 12.88 | 7.73 | $8.79 \times 10^{-30}$ |
| South China Sea | 5.06 | 5.75 | 0.69 | $2.79 \times 10^{-2}$ |

as strong indicators of preference. The result is not merely improved matching, but a rapid narrowing of content diversity and a strong reinforcement of early signals.

This amplification effect is especially evident when comparing broader and more specific topics. Broader themes such as *Russia* and *China* maintained higher relevance throughout the chain, while specific issues like the *Russia–Ukraine War* and *South China Sea Dispute* showed faster drift, particularly in the baseline condition. Even under interest-based viewing, the improvement for narrower topics was much smaller. For example, the South China Sea topic saw only a 0.69 gain in AUC compared to Broader China's 7.73. This indicates that the recommendation algorithm is less responsive to simulated interest for detailed or complex topics.

Moreover, the overall relevance scores remained low. Across all datasets and conditions, average relevance rarely exceeded 0.35 on a 0–2 scale, suggesting that even with simulated interest, the system delivers a substantial volume of marginally related content. Together, these findings point to a form of algorithmic bias rooted in feedback amplification and topic generalization. Minimal behavioral signals are treated as high-confidence preferences, reinforcing certain content paths while allowing others to fade. In short-form environments where interactions are brief and often ambiguous, this dynamic can limit content diversity and reduce exposure to complex or underrepresented narratives.

## 5 Conclusion and Discussion

This study examined how simulated watch-time behavior affects topical relevance in YouTube Shorts recommendations. Using four geopolitical themes and two viewing conditions, one with uniform short views and one with interest-based durations, we evaluated how small differences in watch time shape recommendation paths. Relevance was measured using a generative AI model and quantified through area under the curve (AUC) and paired statistical testing.

Results show that even minor increases in watch time lead to significant changes in recommendation relevance, addressing **RQ1** and **RQ2**. Broader topics like *Russia* and *China* consistently retained higher relevance, while narrower themes such as the *Russia–Ukraine War* and *South China Sea Dispute* showed limited improvement, supporting **RQ3**. This suggests the algorithm amplifies general content and is less responsive to specific or sensitive topics, even when user interest is simulated.

These findings raise concerns about feedback amplification, where minimal signals result in narrowed exposure and reduced content diversity. This has implications for platform design, public discourse, and information access, especially in short-form ecosystems where user signals are brief and easily over-interpreted.

Our work contributes to ongoing discussions across computer science, media studies, and platform governance. It highlights the need for transparent algorithms that account for the social impact of personalization. Future research will expand to other domains and include interactions such as liking, commenting, and scrolling, as well as more fine-grained watch-time variations.

In summary, YouTube Shorts responds disproportionately to subtle behavioral cues, amplifying certain content trajectories while allowing others to diminish. Understanding and addressing these dynamics is essential for building recommendation systems that promote greater fairness, topical diversity, and informational integrity.

**Acknowledgements.** This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Army Research Office (W911NF-23-1-0011, W911NF-24-1-0078, W911NF-25-1-0147), U.S. Office of Naval Research (N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Defense Advanced Research Projects Agency, the Australian Department of Defense Strategic Policy Grants

Program, Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment, and the Donaghey Foundation at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

## References

1. Abul-Fottouh, D., Song, M.Y., Gruzd, A.: Examining algorithmic biases in youtube's recommendations of vaccine videos. International Journal of Medical Informatics **140**, 104175 (2020)
2. BBC: What is the south china sea dispute? (Jul 2023), https://www.bbc.com/news/world-asia-pacific-13748349, accessed: 2025-05-11
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
4. Cakmak, M.C., Agarwal, N.: Unpacking algorithmic bias in youtube shorts by analyzing thumbnails. In: Proceedings of the 58th Hawaii International Conference on System Sciences (January 2025), https://hdl.handle.net/10125/109144
5. Covington, P., Adams, J., Sargin, E.: Deep neural networks for youtube recommendations. In: Proceedings of the 10th ACM conference on recommender systems. pp. 191–198 (2016)
6. Funaiole, M.P., Hart, B.: Unpacking china's naval buildup. Center for Strategic and International Studies (CSIS) (2024), https://www.csis.org/analysis/unpacking-chinas-naval-buildup
7. Fuziondigital: 1.5 billion youtube shorts user engagement monthly (Nov 2021), https://fuziondigital.co.za/our-blog/1-5-billion-youtube-shorts-user-engagement-monthly/, accessed: 2025-05-11
8. Lautman, O.: Putin's nasty soviet nostalgia. Center for European Policy Analysis (CEPA) (2022), https://cepa.org/article/putins-nasty-soviet-nostalgia
9. Masters, J.: Ukraine: Conflict at the crossroads of europe and russia (Feb 2023), https://www.cfr.org/backgrounder/ukraine-conflict-crossroads-europe-and-russia, accessed: 2025-05-11
10. Ng, Y.M.M., Hoffmann Pham, K., Luengo-Oroz, M.: Exploring youtube's recommendation system in the context of covid-19 vaccines: Computational and comparative analysis of video trajectories. Journal of medical Internet research **25**, e49061 (2023)
11. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: Ms marco: A human-generated machine reading comprehension dataset (2016)
12. Parameswaran, P.: What's behind the new china-asean south china sea code of conduct guidelines? Wilson Center (2023), https://www.wilsoncenter.org/blog-post/whats-behind-new-china-asean-south-china-sea-code-conduct-talk-guidelines
13. Streamers: Youtube scraper (2024), https://apify.com/streamers/youtube-scraper, accessed: 2024-01-10
14. The Guardian: Un finds further evidence of russian war crimes in ukraine. The Guardian (2023), https://www.theguardian.com/world/2023/oct/21/un-finds-further-evidence-of-russian-war-crimes-in-ukraine
15. Yang, Y., Yih, W.t., Meek, C.: Wikiqa: A challenge dataset for open-domain question answering. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 2013–2018 (2015)