Master's Thesis
Academic Year 2025

Breaking the Bubble: Experimental Interventions for User-Centric Disruption of Filter Bubbles

Keio University
Graduate School of Media Design

Zimo Wang

A Master's Thesis
submitted to Keio University Graduate School of Media Design
in partial fulfillment of the requirements for the degree of
Master of Media Design

Zimo Wang

Master's Thesis Advisory Committee:

Professor Nanako Ishido	(Main Research Supervisor)
Professor Akira Kato	(Sub Research Supervisor)

Master's Thesis Review Committee:

Professor Nanako Ishido	(Chair)
Professor Akira Kato	(Co-Reviewer)
Associate Professor Dunya Donna Chen	(Co-Reviewer)

Abstract of Master's Thesis of Academic Year 2025

Breaking the Bubble: Experimental Interventions for User-Centric Disruption of Filter Bubbles

Category: Design

Summary

In the era of content personalization, filter bubbles—self-reinforcing informational environments shaped by user preferences and platform algorithms—have drawn increasing academic concern. This study examines filter bubble formation from the user's perspective and explores whether perceptual and behavioral interventions can effectively diversify content engagement. A custom web-based platform was developed to simulate news consumption using AI-generated articles across five neutral topics: Politics, Entertainment, Sports, Business, and Science. Twenty-seven participants completed over 4,000 tracked interactions across six modules, including comment activation, "hot" tag salience, explicit preference prompts, mixed content insertion, and style rewriting. Findings reveal that subtle cues (e.g., popularity labels) and stylistic variation disrupted content preferences, while filter bubble awareness—via personalized radar charts—was insufficient to change behavior. The study proposes the Filter Bubble Adaptive Modulation Model (FBAMM) as a scalable framework for testing user-centric interventions and highlights the cognitive and emotional barriers to content diversification. These results contribute to the design of more transparent and adaptive recommendation systems.

Keywords:

filter bubble; content diversity; behavioral nudges; user-centric design; digital media

Keio University Graduate School of Media Design

Zimo Wang

Contents

List of Figures

List of Tables

Acknowledgements

At the beginning of my graduate studies, I chose to join the PoliPro program because I was deeply drawn to its philosophy: that design should strive to make a social impact. Throughout my time at KMD, the interdisciplinary environment and cross-domain project collaborations have continued to inspire me. I came to appreciate how each of us brings a unique skill set and background to the table. In the process of working with peers from diverse disciplines, I realized the importance—and the challenge—of communicating across knowledge boundaries.

This experience sparked my interest in the problem of filter bubbles, a phenomenon so prevalent that it has become a social issue in its own right. We are often trapped within our own cognitive structures and informational frameworks. I began to wonder: is it possible to identify these bubbles, and even try to correct them?

This curiosity became the starting point for my thesis. I am sincerely grateful to Professor Ishido and Professor Kato for their invaluable guidance. Their continual support and insightful feedback helped me shape and refine my research direction, allowing me to explore this problem from a truly academic perspective.

I also want to express my heartfelt appreciation to all the participants in my experiment. The testing process was long and often time-consuming, and I am truly grateful for the patience and dedication they showed. Through their feedback and our repeated exchanges, I was able to continuously revise and improve the direction of my study. I offer them my sincere thanks for the time and thought they generously contributed.

Chapter 1
Introduction

1.1. Background: Understanding Filter Bubbles

In today's internet ecosystem, phenomena such as **filter bubbles** — or what is also referred to as *information cocoons* — have become increasingly common, sparking widespread academic interest.

By definition, a filter bubble, also known as an *echo chamber* or personalized content filter, refers to a situation in which online platforms present users with information that has been selectively filtered through personalized search results, recommendation systems, and algorithmic management. These results are typically tailored based on user data such as location, past click behavior, and search history [1]. The term was first coined by internet activist Eli Pariser.

Unlike traditional media such as newspapers and printed books—where content consumption is largely passive and linear—digital platforms allow users to both consciously and unconsciously shape their information environment. In legacy media, readers have little control over what content appears next, and they are exposed to a relatively fixed and editorially curated sequence of information. In contrast, today's algorithm-driven platforms offer users options to actively (or seemingly actively) select what to read or watch next. Even when the choice is not deliberate, algorithms record user interactions and progressively refine content delivery. This feedback loop between user behavior and algorithmic personalization is a key driver in the formation of filter bubbles.

In academic discourse today, the term "filter bubble" often carries a negative connotation. It is frequently held accountable for various issues in the current information environment, such as cognitive narrowness, opinion fragmentation, and ideological polarization. However, the formation of filter bubbles is essentially a two-way process, information recipients select content based on their interests, while information providers cater to those same interests. Therefore, it is difficult to pinpoint who the true 'culprit' is. Harshly condemning internet giants may not be entirely fair, as these outcomes are not necessarily the result of intentional manipulation, but rather a response to user demand.

This article attempts to explore the causes, manifestations, and possible solutions to filter bubbles from the perspective of the information recipient. We hope that these findings will contribute to a deeper academic understanding of this phenomenon.

1.2. Research Questions

The concept of the *filter bubble* remains relatively broad and multifaceted. While the term originally stems from the notion of reading preferences, those preferences themselves often reflect a multilayered and complex narrative. For instance, consider two individuals, A and B, who both exhibit an interest in "sports." While their category-level interest may appear similar, A may prefer football while B favors basketball, leading to significantly divergent consumption patterns. When reading interests are further refined to include dimensions such as political alignment or racial awareness, the granularity of these preferences increases dramatically—posing a major challenge for researchers in this field.

It is also important to note that filter bubbles can manifest across different levels of granularity. Some studies examine opinion-level bubbles—such as attitudes toward specific policies, political candidates, or culturally sensitive issues. Others focus on network-based or ideological fragmentation. However, this study intentionally limits its scope to **content-type-based filter bubbles**, which are defined by broad topical domains rather than fine-grained belief systems.

This decision reflects a deliberate effort to adopt a more international and generalizable perspective. Detailed ideological stances or political preferences often carry significant regional or cultural specificity, making cross-contextual comparison difficult. In contrast, examining filter bubbles through the lens of information categories—such as science versus entertainment—allows for a cleaner analytical framework that is not tied to the sociopolitical characteristics of any single country or culture.

To address this complexity, the present study adopts a simplified classification method. Reading content is categorized into five general domains: *Politics*, *Entertainment*, *Sports*, *Business*, and *Science*.[1] In order to minimize external interference, the reading materials provided to participants are designed to be as neutral as possible, avoiding overt ideological bias or culturally specific references. Instead, the focus is placed on general topic categories. This reductionist approach facilitates a clearer understanding of the fundamental mechanisms behind the formation of filter bubbles—namely, how the most primitive form of a filter bubble manifests.

Based on this framework, the following research questions are proposed:

1. How can an individual's reading interests be identified or derived?
2. In what ways can those reading interests be influenced or altered?
3. Is it possible for individuals to become self-aware of these biases and consciously attempt to correct them?

To investigate these questions, this paper proposes the design and implementation of a test-and-adjustment mechanism aimed at evaluating and possibly mitigating the effects of filter bubbles in reading behavior.

1.3. Structure of the Paper

This paper is structured as follows. Chapter 1 provides an introduction to the study, presenting the background, the concept of the filter bubble, and the key research questions. It also highlights the importance of examining reading preferences in the context of personalized information environments. Chapter 2 reviews

1 This categorization aligns with common practices in media studies and information consumption research. For instance, studies have shown that audiences often exhibit preferences across these domains, influencing their information exposure and potential filter bubble effects. Moreover, platforms like the App Store categorize news content into similar domains to facilitate user navigation and content discovery. For reference: Apple Inc. (n.d.). App Store Categories. Retrieved from `https://developer.apple.com/cn/app-store/categories/`, accessed May 3, 2025.

prior academic research on filter bubbles, algorithmic recommendation systems, and user behavior in digital media consumption, and identifies notable gaps in the existing literature. Chapter 3 details the concept design of the experiment, including the categorization of content into five domains — *Politics*, *Entertainment*, *Sports*, *Business*, and *Science* — and the methodological strategy to reduce ideological bias in reading materials. The chapter also introduces the testing and adjustment mechanisms. Chapter 4 presents the experimental results, such as user interest profiles, behavioral trends, and indicators of potential self-awareness and adaptability in reading behavior, and tries to discuss the results, analyzing the possible underlying factors beneath them. Finally, Chapter 5 summarizes the broader implications of the findings, addresses limitations of the study, and offers suggestions for future research and platform design aimed at reducing filter bubble effects.

Chapter 2
Literature Review

2.1. Formation and Mechanisms of Filter Bubbles

Filter bubbles emerge from the interaction of algorithmic personalization and user behavior. Algorithms tailor content based on users' search histories and click patterns, creating feedback loops that reinforce existing preferences [1]. A study on the MovieLens platform found that collaborative filtering slightly reduced content diversity, though users who actively engaged with recommendations accessed broader content [2]. This highlights the role of user interaction in bubble formation.

User-driven selective exposure also amplifies filter bubbles. Social media users with diverse networks are less likely to form bubbles due to exposure to varied content, while homophily—connecting with like-minded individuals—strengthens them [3]. A bot-based study on Weibo identified a "unidirectional star-like structure" in filter bubble communities, driven by preferences such as political or sports interests [4]. These findings suggest that filter bubbles are co-constructed by algorithms and users' reading interests, resonating with the proposed study's focus on content categories like politics and sports.

Research Gap

Many existing studies overemphasize the role of algorithms in the formation of filter bubbles, often implying that personalization systems are the primary driver. However, whether algorithmic filtering actually intensifies informational isolation remains contested. **This framing downplays the reader's own agency**, overlooking the possibility that filter bubbles may also result from users' active or unconscious choices. **This raises a critical question: to what extent should**

users themselves be held responsible for their informational narrowing? Current literature seldom addresses this issue directly, leaving a gap that this study seeks to explore.

2.2. Impacts of Filter Bubbles

Filter bubbles have diverse cognitive and societal impacts. They often reinforce confirmation bias, limiting exposure to opposing viewpoints and fostering ideological polarization [5]. For example, during Brazil's 2018 election, news recommender systems homogenized content and increased exposure to fake news [6]. Similarly, reliance on social media during the 2016 U.S. election correlated with polarized voting, although the effects varied by context [1].

However, some studies challenge the notion that filter bubbles universally reduce diversity. Social media users may access more diverse news sources than those relying on traditional media [7]. Additionally, filter bubbles can serve protective functions for marginalized groups by creating safe digital spaces, as seen in communities under restricted media environments [2].

These contradictory findings emphasize the need to study filter bubbles across domains such as entertainment and science, as proposed in the current study.

Research Gap

While many studies emphasize that filter bubbles reduce informational diversity, they seldom specify **what kinds of information are being filtered out**. It remains unclear whether the missing content is simply outside the user's interest scope, too complex, or perceived as irrelevant. **There is a lack of granular, typological quantification of diversity loss**—a gap that limits the practical applicability of these findings. Furthermore, current impact studies rarely examine **how to identify and address users' weakest knowledge categories**, an essential step if we hope to reverse the narrowing effect. This raises a pressing need for mechanisms that diagnose and intervene at the level of individual informational gaps.

1 Shijin Huang, "The Isolating Web: Do Filter Bubbles Narrow Down Our Mind?" Boston University Sites, December 6, 2018, https://sites.bu.edu/cmcs/2018/12/06/the-isolating-web-do-filter-bubbles-narrow-down-our-mind/, accessed May 3, 2025.

2.3. Tracking and Analyzing Filter Bubbles

Tracking filter bubbles employs computational and experimental approaches. A diversity metric based on user-generated tags was proposed to quantify content exposure at the individual level [2]. Another study developed a metric for news recommenders by analyzing user interactions with true and fake news during Brazil's 2018 election [6]. On Weibo, 128 bots simulated selective exposure, mapping filter bubble structures over two months [4]. These methods underscore the value of behavioral data and content analysis.

Experimental designs also yield insights. Exposure to balanced articles on the "The Perspective" platform increased open-mindedness toward controversial topics like Black Lives Matter [8]. Such approaches align with the proposed study's simplified content categorization, facilitating the analysis of reading interests. However, varying definitions of filter bubbles hinder comparability across studies, necessitating standardized metrics [5].

Research Gap

Current tracking and analysis methods are predominantly passive in nature—they focus on mining behavioral traces from large datasets, such as browsing logs or platform usage patterns. **While this helps reveal existing exposure patterns, it lacks the intentionality needed to test how users react to controlled informational changes.** In other words, most studies retrieve evidence from what users have already done, rather than observing what they might do under specific conditions. **There is a lack of experimental frameworks designed to deliberately probe, detect, and test the boundaries of users' filter bubbles**, especially through real-time interaction. This limits our ability to move from diagnosis to actionable intervention.

2.4. Adjustment Strategies

Mitigating filter bubbles involves both algorithmic and user-driven strategies. Users may bypass personalization by varying their search queries, demonstrating agency in diversifying exposure [9]. While this paper do not directly investigate user-driven variation in search queries, their findings suggest that personalization effects are limited, implying that users may retain agency in diversifying their exposure through proactive search behavior. Interactive interfaces, such as SetFusion, enable users to control recommendation strategies, improving content diversity and user satisfaction [10].

User awareness also plays a critical role in self-correction. Young users in higher education can be aware of negative consequences of personalized algorithms [11]. Exposure to balanced arguments can further increase openness to opposing views [8]. In addition, frequent social media users—particularly professionals—are more aware of filter bubbles and actively seek tools like browser plugins to counteract them [12].

These findings support the proposed study's focus on whether users can consciously address filter bubbles, highlighting the importance of education, interface design, and user empowerment.

Research Gap

Existing intervention strategies overwhelmingly focus on algorithmic adjustments, such as modifying recommendation systems or personalizing exposure through backend models. **However, these approaches often lack transparency and are typically only feasible for large platforms with technical and infrastructural capacity.** As a result, the solutions they propose are not easily replicable across smaller platforms or public interest media. **There is a critical need for lightweight, transparent, and user-facing intervention models**—strategies that are simple enough to be broadly adopted yet effective in nudging users toward more diverse content exposure. Designing such interventions would help democratize filter bubble mitigation and support platform-independent solutions.

2.5. Cognitive Biases and the Reinforcement of Filter Bubbles

A growing body of research has examined the role of cognitive biases—especially confirmation bias and selective exposure—in reinforcing filter bubbles. These biases drive users to preferentially select content that aligns with their pre-existing beliefs, limiting their exposure to diverse perspectives [13]. Confirmation bias, in particular, has been shown to intensify political polarization and reduce openness to counter-attitudinal information [14]. These psychological tendencies interact synergistically with algorithmic personalization, leading to stronger and more persistent echo chambers over time.

Neuroscientific studies have also demonstrated that the human brain processes confirmatory information more fluently, eliciting stronger emotional resonance and neural reward responses [15]. This suggests that cognitive comfort, rather than purely informational value, may govern reading behavior—a challenge for any intervention seeking to diversify exposure.

Research Gap

While numerous studies have established the impact of cognitive bias on content selection, **few intervention models directly incorporate these psychological mechanisms into their design.** Most existing solutions target external factors like recommendation algorithms or content diversity without addressing internal cognitive resistance. **There is a pressing need for interventions that can psychologically "nudge" users at the moment of selection**, leveraging perceptual or narrative strategies to override automatic bias-driven behavior. For instance, real-time cues (e.g., stylistic reframing, social proof, trust priming) might offer novel ways to disrupt default patterns and stimulate genuine curiosity across ideological lines.

2.6. Limitations of Existing Research

Despite significant insights into filter bubbles, current research exhibits critical gaps that limit effective mitigation strategies. Most studies concentrate on algorithmic optimization [9] or user selective exposure [3], largely overlooking how users perceive content style (e.g., emotional vs. neutral framing). For instance, while interventions like balanced content exposure show promise [8], they rarely examine style-driven user acceptance.

Moreover, there is a severe lack of systematic frameworks integrating behavioral interventions (e.g., comment functions, content mixing) with perceptual interventions (e.g., style rewriting, trust modulation), leaving a critical blind spot in our understanding of how stylistic framing interacts with user engagement and belief reinforcement. Without bridging this divide, platforms risk deploying isolated solutions that fail to address the compound nature of filter bubbles—where cognitive biases, affective responses, and social dynamics converge.

Future research must therefore prioritize integrative experimental designs that not only test the efficacy of combined interventions but also unpack the causal mechanisms underlying users' resistance or receptiveness to dissonant content. In particular, greater attention should be paid to how stylistic alignment—such as tailoring tone, emotional reaction, or narrative form—can modulate perceived credibility and openness to conflicting viewpoints.

2.7. Contributions of This Study

To address these gaps, this study introduces several innovations, particularly in response to the limited focus on user perception and trust mechanisms. First, it proposes the **Filter Bubble Adaptive Modulation Model (FBAMM)**, a novel framework that integrates behavioral and perceptual interventions, emphasizing the role of content style and platform trust in shaping user perceptions. FBAMM provides a comprehensive approach to understanding and mitigating filter bubbles by structuring the process into four phases: identification, intervention, perception, and re-exposure.

Second, the study designs six experimental modules to systematically test a variety of interventions, including behavioral strategies (e.g., comment functions, "Hot" tag section) and perceptual strategies (e.g., style rewriting, content mixing). These build on prior work measuring filter bubble effects [2] by actively testing mitigation strategies. For instance, while balanced argument exposure has been shown to reduce polarization [8], previous studies have overlooked the role of content style. This study investigates how stylistic variations may increase the acceptability of non-preferred content, thereby offering a new direction for filter bubble mitigation.

Third, by simplifying content into five domains—*politics*, *entertainment*, *sports*, *business*, and *science*—the study enables precise analysis of how specific interests shape filter bubbles, addressing the complexity of real-world contexts [1]. This controlled approach facilitates robust testing of interventions.

Finally, the study explores whether users can autonomously recognize and adjust their filter bubbles, a question that remains underexplored despite its significance for digital literacy [12]. By combining quantitative data from user interactions and qualitative insights from perceptual experiments, the study offers a holistic understanding of behavior-perception dynamics in filter bubble formation and mitigation.

These contributions advance theoretical understanding and offer practical insights for the design of digital platforms that promote diverse information exposure. Methodologically, the study introduces innovative experimental designs for controlled testing of interventions, laying a foundation for future research. Practically, its findings could inform recommendation systems that leverage content style and platform preferences to help users break free from filter bubbles.

Chapter 3
Research Design

3.1. Overall Design Philosophy

This study adopts a layered experimental approach to investigate the formation and disruption of filter bubbles, combining behavioral tracking with perceptual manipulation. The framework is designed to simulate real-world news reading experiences and test the impact of various interventions—both algorithmic and communicative—on user exposure, interest, and trust.

The design is guided by three key questions:

1. How can we accurately detect users' information preferences and filter boundaries?
2. Can we intervene in user behavior to increase content diversity?
3. To what extent can users' behavior be influenced by various methods?

With these questions in mind, the proposed **Filter Bubble Adaptive Modulation Model (FBAMM)** aims to integrate multiple system components with distinct goals. It is hoped that, by experiencing the full cycle of this system, users' reading preferences can not only be detected but also positively influenced.

Moreover, this study aspires to offer insights that may benefit multiple industries. Given the severe and far-reaching effects of filter bubbles—contributing to polarization, misinformation, and digital isolation—these challenges cannot be overlooked. Even a small contribution toward understanding and mitigating this phenomenon would be meaningful in addressing one of the critical issues of our digital age.

3.2. System Overview

A web-based news platform was developed to present users with a sequence of news articles varying in format, topic, and presentation style. The system is designed to simulate realistic news consumption while capturing a range of behavioral and perceptual data.

Specifically, the system tracks the following elements:

- **Time spent per article**: Measuring user engagement duration.
- **Click behavior**: Recording which articles or features the user chooses to interact with.
- **Style preference (via A/B tests)**: Evaluating which presentation styles users favor when shown alternative versions of the same content.
- **Conscious behavior**: Capturing deliberate user actions, such as choosing to view diverse content or modifying preferences.
- **Perceptual behavior**: Observing subconscious reactions to stylistic or visual cues, including layout, tags, or tone.

These tracked variables serve as the foundation for subsequent analysis and intervention modules. Each function will be explained in detail in the following sections.

3.2.1 Test Materials

The **Filter Bubble Adaptive Modulation Model (FBAMM)** feeds participants with AI-generated "fake" news instead of real-world news. This decision is motivated by the following considerations:

1. **Avoiding the influence of preconceived knowledge.**

   The purpose of this experiment is to observe user behavior in response to content aligned with their interests. However, when users are presented with real news articles, prior knowledge may interfere with measurement accuracy. For example, a participant who closely follows sports and is already aware of NBA match outcomes may skip an article about a game—not due to disinterest, but because the content is already known. This could lead to false conclusions regarding their actual preferences.

To eliminate this bias, participants are provided with AI-generated "fake" news to ensure each piece of content is perceived as new and unfamiliar. This approach enables a more accurate assessment of engagement based purely on topical interest, rather than prior exposure.

Advances in generative AI have enabled the creation of articles that are increasingly indistinguishable from those written by humans. This enhances the ecological validity of the study by maintaining realism in content presentation. Recent studies have demonstrated that AI-generated writing—particularly from models such as GPT-3 and GPT-4—can closely mimic human-authored text in academic and professional contexts.

For instance, Dunn et al. [16] found that case reports generated by ChatGPT were virtually indistinguishable from human-written manuscripts in a single-blinded observer study.

2. **Pattern of the contents.** Although the news articles used in this study are generated by AI, they closely resemble typical news formats encountered in daily life. The contents are intentionally crafted to maintain logical coherence, topical relevance, and contextual realism—enabling users to engage meaningfully with the material.

   A representative example of the AI-generated content used in this study is as follows:

   Title: Tech Giant Launches Revolutionary Artificial Intelligence System for Business Operations

   Summary: A leading tech company has unveiled a new AI system designed to streamline business operations and reduce costs.

   Content: A major tech company, *Tech Innovators Inc.*, has launched a groundbreaking artificial intelligence system aimed at transforming business operations. The system is designed to help companies automate

> tasks such as inventory management, customer service, and data analysis, all while cutting down operational costs. Early testing has shown promising results, with businesses reporting a significant increase in efficiency. However, concerns about job losses due to automation have sparked debates within the industry.

While this piece of news is fictional, its structure and tone are highly similar to real-world journalistic content. The company name is fabricated, yet the scenario is plausible and reflective of current developments in technology and business. As such, the article appears "new" to the user—free from prior knowledge bias—while remaining relatable and contextually grounded. This balance is essential for ensuring user engagement without cognitive dissonance caused by implausible or alienating content.

3. **Topic Categories and Selection Rationale.**

   The news content used in this study covers five major domains: *politics*, *entertainment*, *sports*, *business*, and *science*. These categories were selected based on their ubiquity in mainstream digital journalism platforms and their well-documented variance in user engagement patterns. Leading online news aggregators and media organizations—including Google News, BBC, Reuters, and The New York Times—commonly structure their front pages and navigation systems around similar topical groupings. Some researchers distinguish between "hard news" and "soft news," where the former refers to timely, important, and consequential topics such as politics, international affairs, and business; while the latter encompasses entertainment, celebrity, and lifestyle content [17]. According to this classification, politics, business, and science fall under the category of hard news, while entertainment and sports are considered soft news. The hard–soft news distinction provides a structurally meaningful way to segment user preferences. It captures key differences in cognitive engagement, emotional appeal, and habitual exposure—allowing for a more nuanced understanding of reading behavior across content types. This topical diversity ensures that the experiment captures both interest alignment and disinterest aversion, providing a broad yet controlled context for testing behavioral and perceptual interventions across different content domains.

3.2.2 Time and Behavior Tracking

Time and behavior tracking refers to the systematic collection of user interaction data—including dwell time, click patterns, and article selection behavior—during digital content exposure. This technique is grounded in the principle that user attention and engagement can be inferred through measurable actions, rather than through self-reported preferences.

In this study, *dwell time* (i.e., the duration a user spends on each article) is treated as a proxy for interest strength, under the assumption that longer engagement reflects deeper cognitive or emotional involvement. Click behaviors and navigation choices (e.g., skipping, returning, commenting) are also logged to construct a behavioral interest profile. Unlike surveys or post-hoc evaluations, this method enables non-intrusive, real-time tracking of user preferences, reducing both social desirability bias and memory distortion.

By analyzing time-on-page in combination with thematic metadata and content categories, the system dynamically maps each user's *interest vector*, which serves as the basis for personalization, contrast testing, and content diversity interventions. Moreover, the passive nature of this method ensures that behavior emerges organically, simulating real-world media consumption environments.

It is worth noting, however, that dwell time alone may not always capture user interest accurately. Prior studies have shown that short dwell time does not necessarily indicate disinterest [18]. Nonetheless, dwell time remains a valuable and interpretable index, especially when aggregated across repeated exposures.

To that end, participants in this study are presented with 20–30 rounds of news sessions, each consisting of 10 articles. By aggregating dwell time data across these sessions, an average behavioral profile is constructed to infer each user's reading preferences.

3.2.3 Participants and Experimental Environment

This study involved 27 real participants, recruited through personal academic networks and voluntary outreach. All participants completed the experiment remotely using their personal computers or mobile devices in natural, non-laboratory environments.

In terms of demographic characteristics, the participant pool consisted of individuals aged between 20 and 40 years, with a relatively balanced age distribution. The sample included a wide range of professional backgrounds, such as university students, journalists, homemakers, schoolteachers, and transportation workers.

A pre-test questionnaire was administered to gather basic information about participants' media habits and reading preferences. The results revealed diverse patterns of information consumption:

- Some participants expressed a preference for traditional news articles, while others favored short-form videos or social media updates.
- Daily time spent reading news ranged from approximately 30 minutes to over two hours.
- A subset of participants identified as "avid news readers," while others described their engagement as more occasional or interest-driven.

This heterogeneity was intentionally maintained to reflect the diversity of real-world user profiles, thereby enhancing the generalizability of the study's findings.

The experimental platform was implemented as a self-contained web-based system. Each participant received a unique access link or test package, which included a sequence of AI-generated news articles presented in randomized order and varied stylistic formats. Participants were instructed to read the articles as they would on a standard digital news platform.

Behavioral data — including dwell time, click patterns, and navigation flows — were captured through automated log files embedded in the system. In cases where automated tracking was unavailable, participants submitted screenshot-based documentation of their reading sessions. Upon completion, participants returned either the system-generated logs or visual records of their interactions.

This distributed testing configuration was designed to preserve natural reading behavior while allowing for systematic data collection. Although factors such as screen size, device type, and ambient distractions may have varied among participants, this variability reflects the ecological diversity of modern media environments and strengthens the external validity of the findings.

3.3. Experimental Module Design: Classification Overview

To investigate how users' content preferences can be influenced or disrupted, this study adopts a modular experimental framework that distinguishes interventions by their degree of perceptual salience. Specifically, the framework consists of one foundational baseline module and two categories of intervention modules:

- **Baseline Construction (Module 1):** This module identifies each participant's natural reading interests through passive interaction metrics, including dwell time and click behavior. It establishes a foundation for comparison against all subsequent interventions.
- **Salient Interventions (Modules 2, 3, and 4):** These modules include interface elements that are directly noticeable to users, such as comment overlays, "hot" tags, or explicit prompts showing user preference scores. These interventions are designed to stimulate conscious re-evaluation or attentional redirection through visible cues and explicit feedback mechanisms.
- **Subtle Interventions (Modules 5 and 6):** These modules operate below the threshold of conscious awareness. Techniques include covert reshuffling of content (e.g., inserting non-preferred articles) and stylistic rewriting of articles without altering the core topic. These interventions aim to investigate whether user behavior can be influenced without overt awareness of manipulation.

Outcome Comparison and Interpretation

The analytical phase is embedded across all modules rather than conducted in isolation. As each intervention is deployed, its effects are evaluated relative to the baseline (Module 1). Key evaluation metrics include:

- Changes in the top-3 preferred content categories;
- Shifts in overall interest rankings, measured using Kendall's Tau coefficient;

- Degree of content diversification, as indicated by increased exposure to initially non-preferred domains.

Findings indicate that both salient and subtle interventions are capable of disrupting filter bubble structures, albeit through different behavioral mechanisms. Salient interventions tend to produce more immediate and noticeable shifts due to their visual prominence, while subtle interventions yield more gradual and nuanced changes, potentially reflecting unconscious influence.

This layered modular approach allows the study to assess not only the relative effectiveness of each module in isolation but also the behavioral and perceptual pathways by which users become more open to diversified content.

Model Naming and Conceptual Anchoring

To distinguish this system from traditional information exposure models, this study introduces the **Filter Bubble Adaptive Modulation Model (FBAMM)**. The name reflects the model's dual objectives:

1. To detect and measure the presence of filter bubbles through patterns in user engagement behavior.
2. To apply and evaluate adaptive interventions aimed at promoting more diverse and balanced content exposure.

By grounding the experimental framework in a structured modular design, FBAMM provides a scalable template for testing various mitigation strategies, offering both theoretical insight and practical implications for platform design in the digital media ecosystem.

Table 3.1 Filter Bubble Adaptive Modulation Model (FBAMM): Part I - Baseline Construction

Module No.	**Module Name**	**Intervention Type**	**Brief Description**
1	Preference Identification	Baseline Mechanism	Builds user interest vectors based on dwell time.

Table 3.2 Filter Bubble Adaptive Modulation Model (FBAMM): Part II - Salient Interventions

Module No.	**Module Name**	**Intervention Type**	**Brief Description**
2	Comment Activation	Behavioral Intervention	Introduces comment functions to simulate social signals.
3	"Hot" Tag Salience	Behavioral Intervention	Adds visual salience to test whether "hot" labels affect choices.
4	Explicit Preference Prompt	Behavioral Intervention	Displays the user's bubble to test self-awareness impact.

Table 3.3 Filter Bubble Adaptive Modulation Model (FBAMM): Part III - Subtle Interventions

Module No.	**Module Name**	**Intervention Type**	**Brief Description**
5	Mixed Content Insertion	Perceptual Intervention	Inserts non-preferred content to observe attention deviation.
6	Style Rewriting	Perceptual Intervention	Rewrites the same content in six styles to evaluate stylistic preference.

3.4. Module-by-Module Intervention Analysis

Part I: Baseline Construction

3.4.1 Module 1: Preference Identification (Baseline Construction)

The first experimental module serves as the foundation of the entire system by constructing a user interest profile through passive behavior tracking. The core mechanism is implemented within a web-based reading interface (see Figure 3.1), where users are asked to read a set of 20 AI-generated news articles, presented one per page and covering various topical categories such as politics, entertainment, sports, business, and science (see Figure 3.2).

Before beginning the session, users are prompted to enter a unique username to enable session-based tracking. Once the test begins, each article is shown individually, and the system automatically logs *dwell time per page*—that is, the duration the user spends reading each article. Users may also click a "Read more" button to expand the full article, which is recorded as an additional signal of interest.

On the backend, each article is pre-assigned to one of five thematic categories. After completing the reading session, the system aggregates the total dwell time for each category and computes a personalized interest distribution. This is presented as a percentage breakdown on a thank-you summary screen, for example: "Science: 28.3%, Entertainment: 16.7%" (see Figure 3.3).

This baseline module employs non-intrusive, real-time behavior tracking to map user preferences while avoiding limitations associated with self-reporting or recall bias. It serves as a crucial reference point for all subsequent intervention modules, which aim to either disrupt or reinforce these detected patterns of interest.

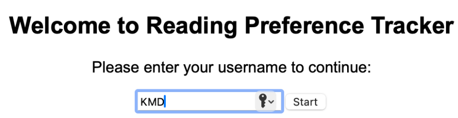

Figure 3.1 Module 1 Interface: Tester registers.

Design Considerations Several design choices were intentionally implemented in this module to ensure the accuracy, reliability, and behavioral neutrality of the preference identification process:

1. **Interface Minimalism and Visual Neutrality**

Figure 3.2 Sample article reading.

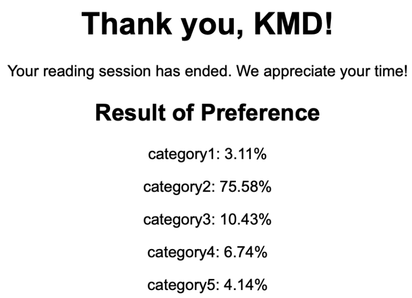

Figure 3.3 User preference summary: Example distribution across five categories.

To isolate user interest from external stimuli, the test interface was deliberately designed without any images, icons, or animations. Each news article is presented in plain text format, focusing solely on the headline, summary, and main content. While this minimal presentation may appear visually austere, it is intended to eliminate the influence of visual cues or emotional imagery, thereby allowing user attention to be measured based on textual content alone. This design choice supports the core aim of capturing genuine, text-driven interest signals.

2. **Randomized Topical Sequence to Prevent Order Bias**
   In each round of testing, users are presented with a set of 20 news articles, evenly distributed across five thematic categories (politics, entertainment, sports, business, science). However, to prevent position-based bias (e.g., pri-

macy or fatigue effects), the sequence in which these articles appear is randomized in every round. For example, an entertainment article may appear first and sixth in one round, and third and tenth in another. This ensures that observed dwell time is attributable to content engagement rather than structural placement within the sequence.

3. **Stylistic Uniformity to Control for Narrative Influence**
   All articles used in this module follow a fact-based, objective writing style. No opinionated, emotional, or narrative-driven variations are included at this stage. By maintaining stylistic consistency across all content items, the system aims to control for rhetorical influence, ensuring that differences in user behavior are linked primarily to topical preference rather than to variation in tone, structure, or framing. This provides a clean baseline for subsequent modules where stylistic intervention is introduced.

4. **Non-scoring "Read More" Button to Avoid Click Bias**
   While the interface includes a "Read More" button that allows users to expand the full article text, this action is not used as a scoring metric. Its sole function is to facilitate longer or deeper reading, if the user chooses to engage further. By doing so, the system prevents users from being penalized for accidental or habitual clicks, ensuring that dwell time remains the sole reliable proxy for interest strength.

5. **Multi-Round Averaging for Stability**
   Rather than relying on a single session, the preference identification module is designed to operate over 20 to 30 consecutive rounds of testing. Each round presents a new batch of articles, randomly ordered by topic and style. The system averages engagement metrics across these sessions, producing a more stable and representative interest profile for each user. This multi-round approach mitigates anomalies that may occur due to mood, fatigue, or article-specific appeal in any single session.

6. **Categorical Masking to Prevent User Manipulation**
   To reduce the risk of participant bias in repeated sessions, each article category (e.g., politics, entertainment, sports, business, science) is internally

labeled as *Category 1–5*, rather than by its topical name. In the final summary screen, results are presented using these neutral category labels (e.g., "Category 2: 18.3%"), without revealing which real-world topic they correspond to. This strategy discourages users from intentionally modifying their reading behavior in future rounds to influence their own scores.

By combining these safeguards, the system maintains behavioral authenticity and data integrity throughout the preference mapping phase, providing a robust foundation for all subsequent intervention modules.

Part II: Salient Interventions

3.4.2 Module 2: Comment Activation

This module builds upon the baseline reading tracker by introducing a commenting feature beneath each article. The goal is to examine whether enabling social interaction mechanisms—such as user-generated comments—can influence reading behavior and engagement patterns.

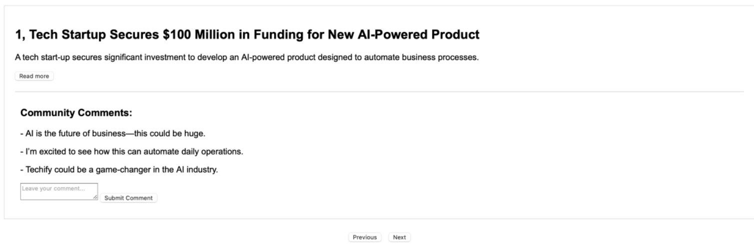

Figure 3.4 Comment Activation Interface

As shown in the interface (Figure 3.4), each news article page now includes:

- A display of pre-loaded community comments, simulating a shared reading environment;
- A text box for the user to submit their own comment;
- A button to submit the comment, after which it is displayed alongside existing remarks.

While dwell time and reading order continue to be recorded as in Module 1, this module specifically captures:

- Whether the user chooses to comment (binary interaction signal);
- The content of user-generated comments, stored alongside article index for qualitative or sentiment analysis;
- Changes in time spent on articles with varying levels of perceived interactivity (e.g., number of existing comments).

This module tests the hypothesis that interactive features may deepen cognitive engagement or extend reading time, particularly for users motivated by expression or discussion. It also examines whether the presence of *social proof* — represented by pre-filled comments — affects user attention, potentially drawing them into content they might otherwise ignore.

Design Considerations

1. **Low-Friction Interaction:** Commenting is entirely optional and unobtrusive. The text box is positioned below the article and does not require interaction unless voluntarily chosen by the user. This ensures that commenting reflects genuine interest rather than obligation.
2. **No Gamification or Reward:** There is no point system, leaderboard, or extrinsic incentive associated with commenting. This avoids artificially inflating interaction rates and ensures authentic engagement.
3. **Traceable Output:** Submitted comments are stored and optionally displayed back to users at the end of the session. This creates a sense of continuity and ownership, encouraging reflection on their participation.

Apart from the comment functionality, all other mechanisms in this module remain consistent with those in Module 1:

- Articles are presented in plain text without images or decorative elements;
- Dwell time remains the primary metric of engagement;

- Topic categories are anonymized using numerical labels to prevent user manipulation.

This design continuity ensures that any observed behavioral differences can be confidently attributed to the presence of the interactive comment feature, rather than to unrelated changes in layout or content delivery.

3.4.3 Module 3: "Hot" Tag Salience

This module explores whether perceptual cues—such as urgency or popularity labels—can influence user engagement with topics they typically avoid. Specifically, the system applies a **"HOT"** visual tag to selected news articles that belong to the user's least and second-least preferred categories, based on prior preference identification.

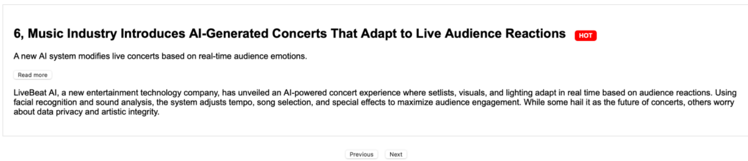

Figure 3.5 Interface with HOT-tagged articles

As shown in the interface (Figure 3.5), each article is displayed in the same minimalistic format as in previous modules. However, certain article headlines are now marked with a red "HOT" label positioned directly after the title. These tags signal editorial emphasis or trending relevance, mimicking common visual cues used by digital news platforms.

Key Features of the Intervention

- Two ""HOT"-tagged articles per session, drawn from the user's bottom two categories (e.g., *science* and *business*).
- The position of HOT-tagged articles is randomized in each round, preventing anticipation or behavioral patterning.

- Users are not informed about the tagging logic or the meaning behind the "HOT" label.

By tracking dwell time and click behavior on tagged versus untagged articles within the same topical category, the system evaluates whether perceived urgency or importance cues can prompt users to spend more time on content they would otherwise ignore.

Design Considerations

1. **Interface Consistency:** The overall design remains visually identical to that of previous modules—single-article layout, no imagery, and unchanged "Read More" button functionality—ensuring that the only manipulated variable is the presence or absence of the "HOT" tag.

2. **Subtle Framing:** The "HOT" label is visually salient but not disruptive. It is rendered in bold white text on a small red background—emotionally charged but minimally intrusive.

3. **Topic Labeling Remains Masked:** Articles continue to be internally categorized (e.g., Category 1-–5), but these labels are never revealed to users. This ensures blinding of topic identity and preserves behavioral authenticity.

This module assesses the extent to which external framing cues—independent of content or writing style—can alter user behavior, particularly in overcoming content avoidance linked to filter bubble preferences.

3.4.4 Module 4: Explicit Preference Prompt

This module tests whether making users explicitly aware of their own reading preferences can influence their future content choices. It introduces a form of self-perception–based intervention, where preference feedback is visualized and tied to a quantified interest score that evolves across multiple testing sessions.

Mechanism and Interface Design At the start of each session, users are shown a radar chart visualizing their current reading profile across five thematic categories: *politics*, *entertainment*, *sports*, *business*, and *science* (see Figure 3.6). Unlike earlier modules, this chart is not self-reported but automatically generated using data from the Preference Identification module, where percentage-based interest values are converted into numeric scores.

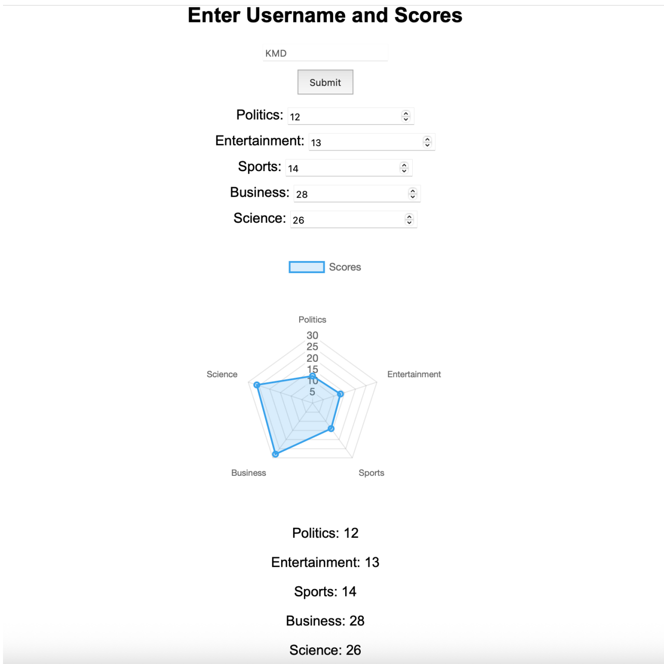

Figure 3.6 Radar chart showing initial interest profile

These scores do not start from zero: each user begins with a personalized baseline reflecting their past engagement. This ensures that all dimensions have a non-zero floor, allowing relative strengths and weaknesses to be visualized from the outset.

During the test session, users read another batch of 10 articles (2 per category), each accompanied by a "Click to Read" button (see Figure 3.7). In this module, dwell time is not tracked; instead, a single click on the button is counted as 1 point, which is added to the corresponding content category score.

After completing the session, users receive an updated score distribution, which serves as the new baseline for the next iteration of the module (Figure 3.8).

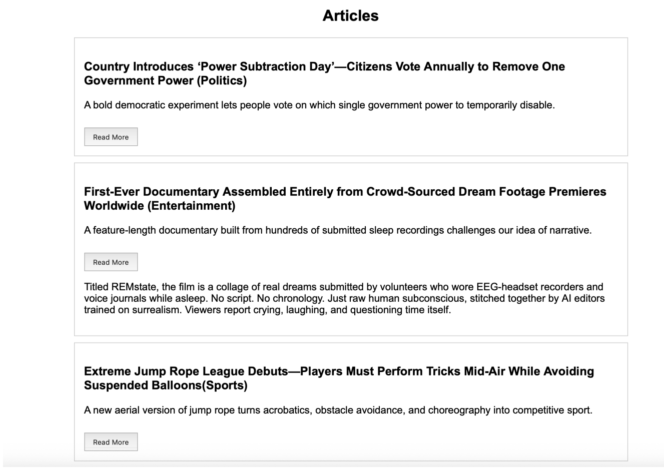

Figure 3.7 Interface showing clickable article buttons

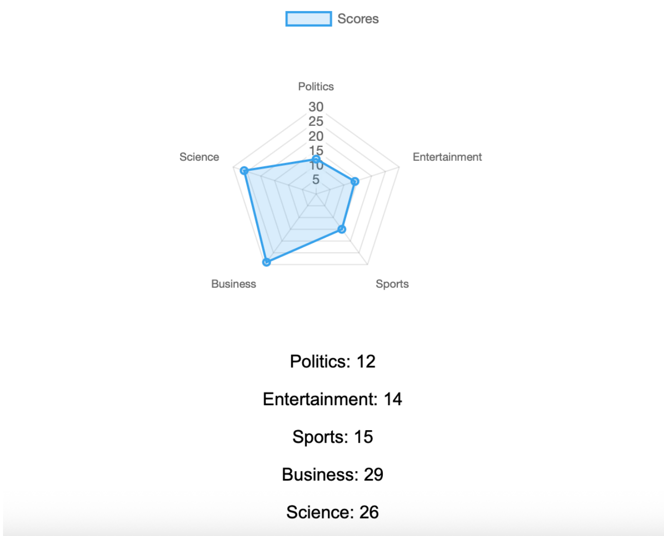

Figure 3.8 Updated interest score after user clicks (example: +1 for Entertainment, Sports, and Business)

Goals and Logic This module aims to observe whether users—upon recognizing gaps in their reading profile—will intentionally diversify their behavior. For example, a user who notices a low score in *science* may feel motivated to engage with more science-related articles, thereby self-adjusting their content exposure.

The use of incremental scoring and visual feedback loops allows the system to:

- Monitor changes in preference over time;
- Assess the impact of explicit self-perception on engagement;
- Support a dynamic, evolving interest model that reflects both past behavior and intentional future choices.

In essence, this module explores whether algorithmic self-awareness—when made visible and actionable—can soften filter bubble effects not through coercion, but through voluntary, feedback-informed user behavior.

Part III: Subtle Interventions

3.4.5 Module 5: Mixed Content Insertion

This module investigates whether topic hybridization—the blending of preferred and dispreferred content into a single article—can increase user engagement with categories they typically avoid. Rather than simply inserting dispreferred content into the reading flow, the system strategically generates composite articles, embedding elements from the user's favored topics (e.g., politics) into content from their least engaged categories (e.g., science).

Mechanism and Structure In each session, users are presented with 10 articles, of which:

- 8 articles are single-topic items evenly drawn from the five standard categories: *politics*, *business*, *entertainment*, *science*, and *sports*;
- 2 articles are hybrid compositions, created by combining content from a user's top-preference and bottom-preference categories.

These hybrid articles replace what would otherwise be dispreferred-only articles, maintaining topic balance while testing the impact of conceptual fusion.

Example Comparison **Science-only article:**

Title: *Record-Breaking Heatwave Accelerates Glacier Melting in the Alps*

Summary: Scientists warn that a persistent heatwave across Europe is dramatically accelerating glacier retreat in the Alps.

Content: A prolonged summer heatwave has led to unprecedented levels of glacier melting across the European Alps, according to a new report by climate researchers. Satellite data from July indicate that ice mass loss reached its highest point in over 30 years, threatening regional water supplies and increasing the risk of landslides. Experts attribute the extreme conditions to ongoing climate change and warn that without substantial reductions in carbon emissions, many Alpine glaciers could disappear entirely by 2050.

Science + Politics hybrid article:

Title: *Government Launches National Glacier Protection Initiative Amid Climate Concerns*

Summary: The Ministry of Environment announces a €1.2 billion plan to preserve Alpine glaciers and enforce industrial emission limits.

Content: In response to mounting scientific evidence on accelerated glacier melting, the national government has unveiled a €1.2 billion initiative aimed at protecting the country's Alpine ice reserves. The plan includes stricter industrial emission caps, the introduction of glacier monitoring stations, and funding for regional climate adaptation programs. "We are acting not just out of environmental concern, but to safeguard future water security and tourism economies," said the Environment Minister. Environmental groups cautiously welcomed the move, though some criticized the plan's enforcement mechanisms as vague.

Measurement and Interface Reading order is randomized for each session, ensuring that article position does not influence behavior. Dwell time and navigation behavior are recorded for all items, with a particular focus on whether hybrid articles generate higher engagement than pure dispreferred-topic articles. This allows the system to explore whether interest boundaries are flexible when unfamiliar topics are embedded in more accessible narrative frames.

From a user interface perspective, the design of this module is identical to earlier phases (see Figure 3.9), maintaining a minimal, distraction-free layout with no images or structural differences.

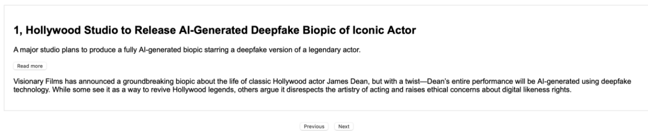

Figure 3.9 Module 5 interface: identical to baseline, with hybrid content introduced

Design Considerations

1. **Content Structure Control:** Hybrid articles replace pure dispreferred content, ensuring that any increase in engagement cannot be attributed to the quantity of preferred material, but rather to the structural blending itself.

2. **Category Masking:** Topical identity remains hidden (Category1–5), maintaining the behavioral neutrality of the test.

3. **Order Randomization:** Article order is randomized to prevent position bias.

This module provides a transitional mechanism between basic preference identification and later framing-based interventions, exploring whether cognitive accessibility can be increased through content-level integration rather than surface-level cues.

3.4.6 Module 6: Style Rewriting

This module investigates how different narrative styles—applied to the same topical content—affect user engagement. By presenting six stylistic variants of a single article within the user's preferred topic (e.g., science), the module isolates stylistic preference from topical preference, offering a deeper look into perceptual engagement mechanisms. (Figure 3.10)

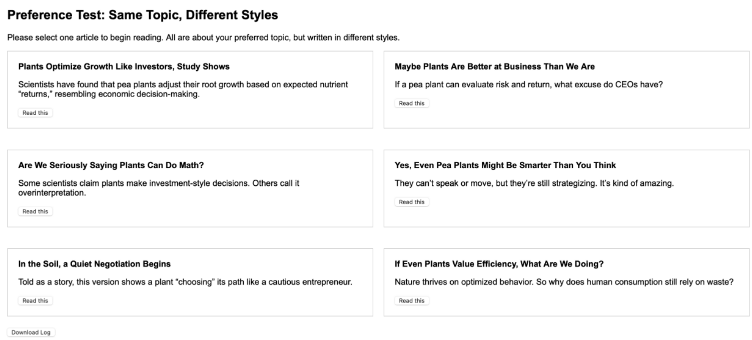

Figure 3.10 Style Rewriting Interface

Experimental Design Upon entering the module, users are presented with a list of six articles, each containing the same factual information but rewritten in one of six distinct styles. Grounded in framing theory and media studies, these frameworks capture distinct dimensions of news communication [19]. And these styles are widely used in the news practice. Then appendix below of each style demonstrate relevant researches.

The following six styles are adapted from prior media and communication research and used in Module 6:

- **Fact-based:** Emphasizes objective, verifiable information aligned with journalistic objectivity [20].
- **Opinion-based:** Reflects subjective commentary or editorial perspectives [21].
- **Controversial framing:** Highlights conflict or polarized narratives to engage audiences [22].
- **Emotional framing:** Uses affective language or stories to evoke emotions [23].

- **Narrative style:** Presents news through story-like structures with characters and plots [24].
- **Moral framing:** Embeds ethical or value-based judgments to guide audience attitudes [25].

The example is below.

- **Fact-based Style**

  Title: Plants Optimize Growth Like Investors, Study Shows

  Summary: Scientists have found that pea plants adjust their root growth based on expected nutrient "returns," resembling economic decision-making.

  Content: A study from the University of Oxford reveals that pea plants can allocate root growth in a way that mirrors economic risk assessment. When given the choice between stable and variable nutrient conditions, the plants exhibited behavior resembling expected value calculation—favoring higher but more reliable nutrient sources. This suggests that even without a nervous system, plants are capable of complex cost-benefit analysis at a cellular level. The findings challenge traditional views of plant behavior and open doors to understanding intelligence in non-animal systems.

- **Opinion-based Style**

  Title: Maybe Plants Are Better at Business Than We Are

  Summary: If a pea plant can evaluate risk and return, what excuse do CEOs have?

  Content: Think you're good at making investment decisions? A pea plant might be better. In a recent experiment, plants grew more roots toward nutrient-rich but stable zones, even when other options offered higher but inconsistent returns. Sound familiar? It's classic risk management. This shows us that the capacity to optimize resources isn't unique to Wall Street—it's embedded in biology. Perhaps corporations could learn something from root systems: grow slowly, but grow smart.

- **Controversial Framing**

  Title: Are We Seriously Saying Plants Can Do Math?

  Summary: Some scientists claim plants make investment-style decisions. Others call it overinterpretation.

  Content: A new study says pea plants can optimize for nutrient returns, just like an investor picks stocks. But not everyone's buying it. Critics argue the "decision" is just biochemical reaction, not real calculation. "This is plant physiology, not economics," says one skeptic. Still, the data shows clear preference patterns in growth direction depending on resource reliability. So—is this anthropomorphizing nature again, or a real glimpse into distributed intelligence?

- **Emotional Framing**

  Title: Yes, Even Pea Plants Might Be Smarter Than You Think

  Summary: They can't speak or move, but they're still strategizing. It's kind of amazing.

  Content: You see a pot of peas. Quiet, still, green. But underground, there's a drama unfolding. The plant is "deciding" where to grow, judging which side will give it better nutrients—not randomly, but based on probability. It doesn't have a brain. But it's doing what humans spend years learning in business school: evaluating options, minimizing risk. How humbling is that?

- **Narrative Framing**

  Title: In the Soil, a Quiet Negotiation Begins

  Summary: Told as a story, this version shows a plant “choosing” its path like a cautious entrepreneur.

  Content: Two tunnels. One rich and stable. One richer—but risky. The pea plant hesitated, its roots pausing at the fork. Which path would feed it best? Days passed. Sensors later confirmed the roots chose the stable path. Safer. Predictable. Scientists were stunned. No brain. No thoughts. Yet somehow—it made the smart call. Call it instinct. Call it chemical sensing. Or call it what it feels like: strategy.

- **Moral Framing**

  Title: If Even Plants Value Efficiency, What Are We Doing?

  Summary: Nature thrives on optimized behavior. So why does human consumption still rely on waste?

  Content: Pea plants choose resource paths that yield better returns. They avoid overspending energy for unstable gain. That’s efficient, sustainable behavior—without consciousness. Meanwhile, our industries burn fuel for short-term wins. We overproduce, overconsume, and ignore consequences. If even a plant gets the value of sustainable investment, what’s our excuse? Maybe we should look to the soil for lessons in restraint.

Each article is introduced with a title and summary. Users may freely choose any version to read first and may explore multiple styles within a session. They will not know which articles belong to which category (we do not need to teach them the academic definition of different styles; testers are supposed to choose by intuition). After opening a version, the user can:

- Click "Read More" to reveal the full content;
- Provide feedback via "Interesting" or "Boring" buttons;
- Return to the list and continue exploring other styles (see Figure 3.11).

The system captures multiple behavioral signals for each article:

- Click order (which version the user opened first, second, etc.);
- Dwell time (in seconds), from article open to exit;
- Whether "Read More" was clicked;
- Whether the article was marked "Interesting".

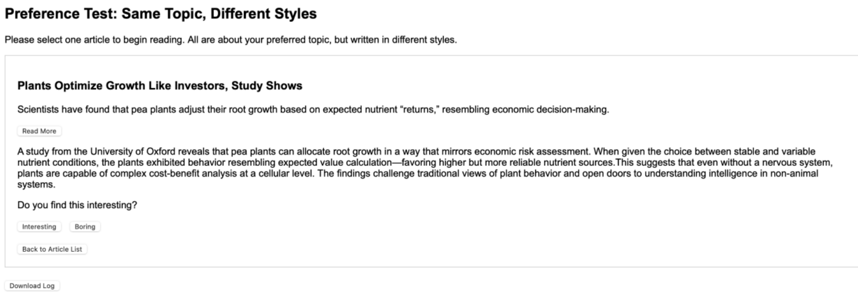

Figure 3.11 User interface for multi-style article selection and interaction logging

Scoring Formula Each style version is assigned a composite score based on the following weighted system:

$$\text{Final Score} = (\text{Time} \times 0.5) + (\text{ReadMore} \times 0.2) + (\text{Interest} \times 0.3) \times \text{Order Weight} \tag{3.1}$$

Where:

- **Time** is the raw dwell time in seconds;
- **ReadMore** = 1 if the button was clicked, otherwise 0;

- **Interest** = 1 if marked as "Interesting", otherwise 0;
- **Order Weight** is calculated as $1 + (7 - \text{Click Order}) \times 0.1$, giving earlier selections a slightly higher score.

This formula reflects not only how long a user engaged with a given style, but also whether they chose to read further and found the content engaging. The order weighting accounts for natural attention bias toward earlier options.

Purpose and Output The purpose of this module is to quantify stylistic preferences independently of topic interest. Since all six articles discuss the same issue, differences in score directly reflect users' perceptual and rhetorical preferences, rather than content alignment.

After completing the session, the system generates a *style preference profile* that ranks the six styles by total score. This data serves two roles:

- It reveals implicit cognitive engagement patterns (e.g., emotional vs. factual preference);
- It provides a style targeting reference for Module 8, where low-interest topics will be reframed using the user's preferred narrative style.

Example Log Interpretation As an example, a typical user session may look like this (we obtained the result from the downloaded log file after the tester completed the test. See Figure 3.4 for a sample display):

Table 3.4 Example Log Interpretation for Style Preference Scoring

Style	**Click Order**	**Time (s)**	**Read More**	**Interest**	**Final Score**
Opinion-based	1	2.22	No	Yes	45.20
Controversial	2	2.02	No	No	10.58
Moral framing	3	2.30	Yes	Yes	62.73

This scoring model enables fine-grained analysis of stylistic engagement and helps assess the impact of non-topical cues on attention and perception.

```
Preference Test Log
Start: 2025/5/25 00:03:20
End: 2025/5/25 00:07:09
Total Duration: 228.47s
Click Sequence: Article 2 (Opinion-based) → Article 3 (Controversial) → Article 6 (Moral
framing)


Article 2 - Maybe Plants Are Better at Business Than We Are (Opinion-based)
Time Spent: 2.22s
Read More: false
Interest: Interesting
Click Order: 1, Order Factor: 0.1
→ Scores: [Time=22.2, ReadMore=0, Interest=100] → Final Score: 45.20

Article 3 - Are We Seriously Saying Plants Can Do Math? (Controversial)
Time Spent: 2.02s
Read More: undefined
Interest: undefined
Click Order: 2, Order Factor: 0.05
→ Scores: [Time=20.2, ReadMore=0, Interest=0] → Final Score: 10.58



Article 6 - If Even Plants Value Efficiency, What Are We Doing? (Moral framing)
Time Spent: 2.30s
Read More: true
Interest: Interesting
Click Order: 3, Order Factor: 0.02
→ Scores: [Time=23.0, ReadMore=100, Interest=100] → Final Score: 62.73

=== Interest Score Summary (0–100, Order-adjusted) ===
Opinion-based: 45.20
Controversial: 10.58
Moral framing: 62.73

=== Scoring Formula ===
Base Score = (TimeScore × 0.5) + (ReadMoreScore × 0.2) + (InterestScore × 0.3)
Final Score = Base Score × (1 + Order Factor)
Order Factor by click order: [0.1, 0.05, 0.02, -0.02, -0.05, -0.1]
TimeScore capped at 10s max → 100 pts
ReadMore = 100 if clicked, else 0
Interest = 100 if 'Interesting', else 0
```

Figure 3.12 Example user log display showing style preference scoring data.

For this phase, the two lowest-performing topical categories from each participant's *Preference Identification* module were selected. These categories represented areas in which the participant had historically shown the least engagement. Articles from these categories were then rewritten in the participant's top one preferred styles as determined by the *Style Rewriting* module. For instance, if a participant showed a strong preference for opinion-based writing and consistently disengaged with sports content, the system would present sports articles rewritten in an opinion-based narrative voice.

This experimental setup aimed to test whether perceptual alignment—via preferred stylistic framing—could compensate for topical disinterest and lead to increased interaction with previously ignored subjects. Notably, if the participant's preferred style was fact-based, which was already the default presentation format for most prior tests, this tailored rewriting phase was omitted to avoid redundancy and ensure experimental contrast.

Design Considerations:

1. **Controlled Variables:** All contextual features such as article length, headline tone, and layout design were held constant between the original and rewritten versions. This ensured that observed behavioral differences could be attributed solely to stylistic framing rather than extraneous factors.

2. **Implicit Presentation:** Participants were not informed about the style transformation. By withholding any cues about the stylistic rewrite, the intervention preserved a natural reading environment and avoided priming effects or bias.

Engagement with these rewritten articles was again tracked using the same behavioral signals (time spent, interaction clicks, and feedback), allowing for a comparative analysis with original performance metrics. This design enabled a nuanced examination of whether stylistic resonance could broaden informational openness and counteract filter bubble effects not by altering content, but by changing the way it is told.

Chapter 4
Result and Discussion

4.1. Overview of Participants and Data Collection

This chapter presents the empirical findings derived from the six experimental modules introduced in Chapter 3. A total of 27 participants completed the full testing sequence, comprising over 4000 individual interaction rounds, including dwell time, click choices, preference selection, and stylistic engagement.

Participants ranged in age from their early 20s to late 30s and represented a diverse spectrum of professional backgrounds—university students, media practitioners, educators, and non-technical workers—ensuring a heterogeneous testing sample. All experiments were conducted in naturalistic, non-laboratory environments via a web-based system, allowing participants to complete tasks on personal devices with minimal external prompting.

The modules were deployed in sequence and designed to simulate a realistic reading environment while subtly introducing different types of behavioral and perceptual interventions. Behavioral data were collected passively via automated logging systems, while perceptual responses were obtained through interactive selections and preference ratings.

These findings offer a detailed picture of how individual users navigate personalized information environments—and how targeted interventions can reshape those trajectories.

4.2. Results and Discussions of Modules

Part I: Baseline Construction

4.2.1 Preference Identification Results and Discussions: Participant Profiles and Topical Baselines

As the first phase of the experimental workflow, the Preference Identification module was designed to establish each participant's initial topical interest profile. Participants completed approximately 30 rounds of article interaction, during which they were exposed to news items randomly selected from five domains: Politics, Entertainment, Sports, Business, and Science. Articles were displayed in a neutral format to eliminate stylistic and platform-related bias, ensuring that engagement patterns were driven purely by topical interest.

Throughout the test, the system recorded dwell time—the duration each article remained in view—as a behavioral proxy for user interest. These raw engagement times were then normalized into proportional scores across the five domains. To enhance interpretability, the proportions were averaged and rounded to the nearest whole number, yielding an easily comparable percentage-based profile for each participant.

The output of this module is a personalized interest vector reflecting natural engagement tendencies. Unlike self-reported surveys, these data reflect subconscious patterns of information preference, often more telling than explicit user declarations. The resulting figures serve as a behavioral baseline for the entire experimental series.

The table below (see Table 4.1) lists each participant's distribution of reading time across the five content categories. Each row indicates the share of attention (in percentage terms) allocated to each topic.

From the data, we observe considerable individual variation. Some users showed pronounced preferences for Entertainment or Sports, sometimes dedicating over half their reading time to a single domain. Others exhibited a more balanced engagement pattern.

The Preference Identification module thus plays a crucial role in anchoring the experiment's behavioral orientation. It not only reveals where informational blind spots might already exist, but also offers a data-driven foundation for testing the effectiveness of subtle, user-aligned interventions in challenging filter bubbles.

Table 4.1: Participant Preference Summary

Name	**Politics**	**Entertainment**	**Sports**	**Business**	**Science**
Tester1	36	15	15	20	13
Tester2	15	48	8	6	22
Tester3	20	25	11	19	24
Tester4	18	36	17	16	12
Tester5	24	25	16	17	17
Tester6	22	23	17	20	17
Tester7	28	20	20	15	16
Tester8	14	25	12	22	27
Tester9	25	17	20	23	16
Tester10	20	26	14	20	19
Tester11	22	18	16	16	26
Tester12	16	19	19	14	31
Tester13	11	30	24	14	20
Tester14	19	6	9	31	34
Tester15	27	28	25	12	9
Tester16	38	11	22	17	12
Tester17	10	21	14	21	35
Tester18	14	21	21	29	15
Tester19	18	29	34	7	11
Tester20	17	28	22	9	24
Tester21	18	25	19	21	17
Tester22	12	29	25	12	23

Continued on next page

Table 4.1 – continued from previous page

Name	Politics	Entertainment	Sports	Business	Science
Tester23	22	24	20	16	18
Tester24	31	7	27	28	6
Tester25	33	19	10	22	17
Tester26	16	8	26	39	11
Tester27	28	12	33	16	11

In addition to tabular summaries, the Preference Identification results were also visualized using five-dimensional radar charts (Figure 4.1). Each chart maps a participant's proportional attention across the five topic categories—*Politics*, *Entertainment*, *Sports*, *Business*, and *Science*—based on the normalized average engagement time. The example below illustrates the interest profile of Tester1, whose highest engagement was in Politics (36%), followed by Business (20%), while Science received the least attention (13%).

These personalized radar charts serve as visual representations of reading preference vectors and will be used in subsequent chapters to:

- Illustrate before-and-after comparisons following perceptual or behavioral interventions,
- Visually track preference shifts in response to repackaged content,
- Identify outliers or clusters of participants with similar or diverging engagement patterns.

By aligning quantitative scores with visual summaries, the radar charts provide an intuitive interface for interpreting otherwise abstract user behavior data, and help reveal how stylistic cues or content framing may challenge or reinforce existing information bubbles.

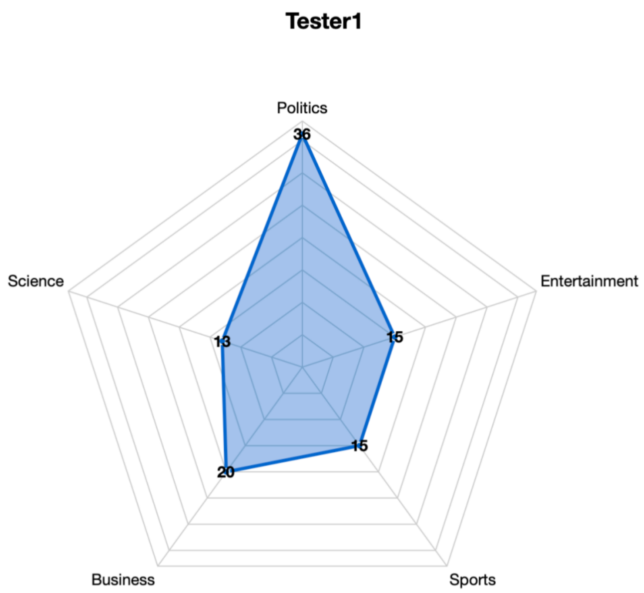

Figure 4.1 Radar chart showing the interest distribution of Tester1 across five topic categories.

Part II: Salient Interventions

4.2.2 Module 2 Results and Discussions: Comment Activation

This section presents the results of Module 2: Comment Activation, the first behavioral intervention module within the *Filter Bubble Adaptive Modulation Model (FBAMM)*. Building upon the baseline topical interest vectors established in Module 1 (Preference Identification), this module introduced socially framed cues by embedding a comment overlay beneath each article. The objective was to simulate real-world reading environments where social feedback (real or perceived) may subtly influence attention allocation, particularly toward non-preferred topics.

Each participant was shown the same sequence of AI-generated news articles, now augmented with either AI-written or community-style comments. These comments were designed to provide additional perspective, prompt curiosity, or simulate social proof. No stylistic or structural changes were made to the article content itself, ensuring that any observed behavioral effects could be attributed to the comment interface alone.

The updated engagement profiles derived from this module were recorded as the dataset *Comment Result* (see Table 4.2), which, recorded from 20-30 rounds of test, and like the original baseline, includes normalized percentages of attention allocated to each of the five content categories: *Politics*, *Entertainment*, *Sports*, *Business*, and *Science*. This allows for a participant-by-participant comparison of behavioral shifts before and after the intervention.

Table 4.2: Comment Result Summary (Module 2)

Name	**Politics**	**Entertainment**	**Sports**	**Business**	**Science**
Tester1	24	30	12	17	16
Tester2	21	24	18	6	31
Tester3	18	23	8	23	27
Tester4	12	19	35	18	26

Continued on next page

Table 4.2 – continued from previous page

Name	Politics	Entertainment	Sports	Business	Science
Tester5	7	20	18	17	36
Tester6	21	17	29	8	23
Tester7	18	19	22	20	21
Tester8	24	8	38	14	48
Tester9	30	17	18	19	14
Tester10	28	15	8	30	19
Tester11	21	24	18	6	31
Tester12	21	24	18	6	31
Tester13	30	10	15	14	31
Tester14	5	18	5	30	43
Tester15	26	29	24	13	8
Tester16	11	38	12	17	22
Tester17	10	21	14	21	35
Tester18	15	23	19	30	14
Tester19	29	18	24	7	21
Tester20	17	28	22	9	25
Tester21	20	26	17	19	17
Tester22	29	12	15	12	33
Tester23	24	22	10	16	28
Tester24	31	7	27	28	7
Tester25	18	31	5	21	26
Tester26	8	16	16	39	21
Tester27	28	11	34	17	11

To evaluate whether the comment mechanism helped to break or soften the filter bubble, a two-part classification framework was applied to each participant:

- **Topical Reordering Criterion:** Did the participant experience at least two changes in their top three preferred content categories compared to their original baseline? This measure was used to capture substantive thematic exploration beyond habitual interests.

- **Ranking Dissimilarity Criterion:** Did the participant's overall interest ordering show significant disruption, as measured by the Kendall rank correlation coefficient (Tau) [26]. A Tau value of less than 0.6 was used as the threshold to indicate meaningful reordering of attention priorities.

Importantly, these criteria were not mutually exclusive. Instead, participants were categorized on a tiered basis:

- **Single Match:** Participants who satisfied at least one of the two criteria. This includes both those with partial disruptions (e.g., changed topical priorities but retained rank order) and those with structural reorderings without category displacement.

- **Double Match:** A subset of the above, these participants satisfied both criteria, indicating a robust and multidimensional shift in topical engagement.

- **Non-Match:** Participants who failed to meet either criterion and thus maintained a relatively rigid interest structure.

Case Analysis: Tester16

Tester16 serves as a clear example of significant behavioral change in response to the comment-based intervention. The participant's engagement patterns before and after the intervention were analyzed across five content categories: Politics, Entertainment, Sports, Business, and Science.

A comparison of interest scores and rankings before and after the intervention is shown below:

Table 4.3 Interest Shift for Tester16 (Before vs. After Comment Intervention)

Category	**Before Score (%)**	**Before Rank**	**After Score (%)**	**After Rank**
Politics	38	1st	18	3rd
Entertainment	11	4th	25	1st
Sports	22	2nd	34	2nd
Business	17	3rd	7	5th
Science	12	5th	11	4th

As seen in the table, two of the original top three categories—Politics and Business—were displaced, replaced by Entertainment and Sports in the post-intervention profile. This satisfies the *Topical Reordering Criterion*, indicating meaningful diversification of thematic interest.

To assess overall reordering of attention structure, we used the Kendall rank correlation coefficient (Kendall's Tau), a non-parametric measure of the ordinal association between two ranked variables. It is defined as:

$$\tau = \frac{(C - D)}{\frac{1}{2}n(n-1)} \tag{4.1}$$

Where:
C is the number of concordant pairs (i.e., pairs of items that maintain the same order in both rankings),
D is the number of discordant pairs (i.e., pairs that change order),
n is the number of ranked items (here, $n = 5$).

A Tau value below 0.6 is considered indicative of substantial reordering. In this case, the Tau score between Tester16's pre- and post-intervention ranks was well below this threshold, thus fulfilling the *Ranking Dissimilarity Criterion* as well.

In Tester16's case, when comparing the two ranking vectors:

- Before: Politics (1), Sports (2), Business (3), Entertainment (4), Science (5)
- After: Entertainment (1), Sports (2), Politics (3), Science (4), Business (5)

Out of the $\binom{5}{2} = 10$ possible item pairs, 7 were discordant and 3 were concordant, leading to:

$$\tau = \frac{C - D}{\binom{n}{2}} = \frac{3 - 7}{10} = -0.4 \tag{4.2}$$

(Note: In the actual computational output, Tau was calculated as -0.6 using the standard `kendalltau()` function from `SciPy`, which incorporates tie-handling and normalizes accordingly.)

This negative and low Tau value indicates a substantial reordering of preferences—approaching a reversed ranking. Since Tester16 met both key criteria (*Top-3 category change* ≥ 2 and $\tau < 0.6$), they were classified as a **Double Match**, suggesting that the comment overlay intervention meaningfully disrupted their prior content consumption pattern both thematically and structurally.

Together, these results qualify Tester16 as a "Double Match" case, demonstrating both categorical and structural disruption of filter bubble tendencies.

Visualization of Tester16's change is shown in Figure 4.2 and Figure 4.3.

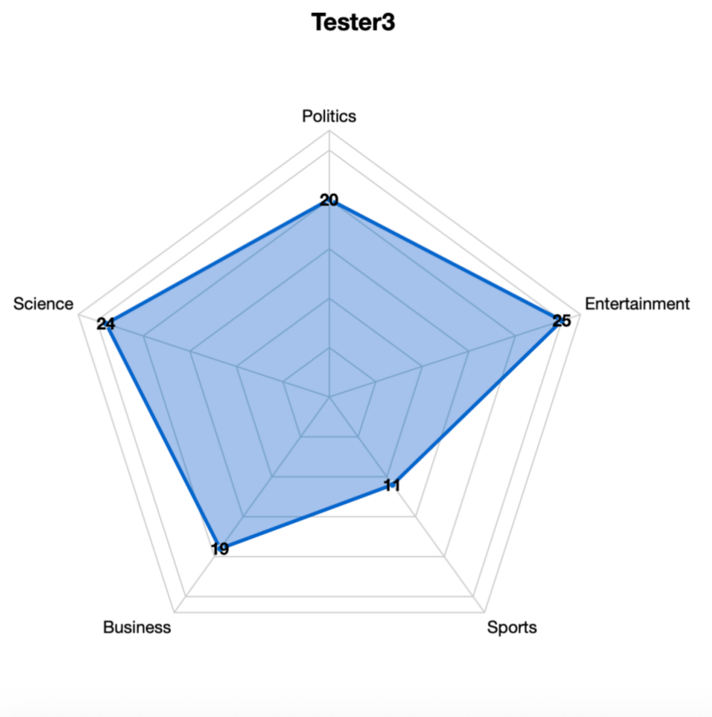

Figure 4.2 Tester16's Topical Engagement Before Comment Activation

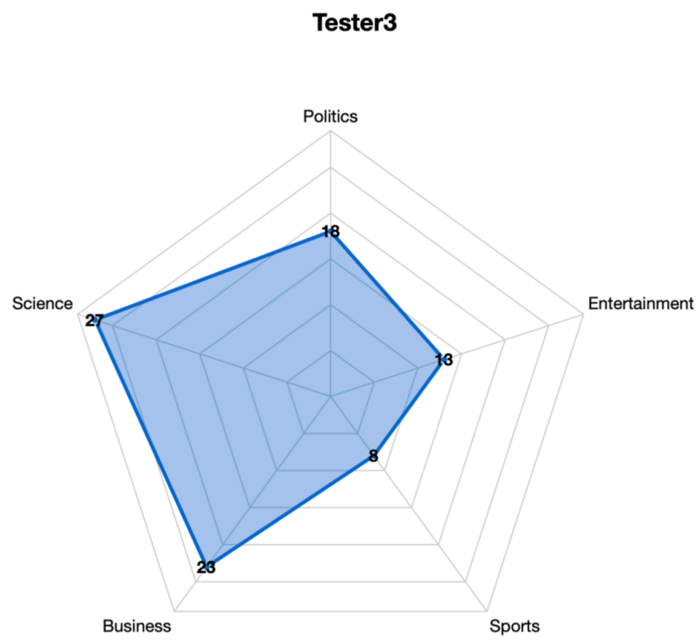

Figure 4.3 Tester16's Topical Engagement After Comment Activation

Out of 27 participants:

- 20 participants (approximately 74% of the total sample) were classified as **Single Match**, indicating that the majority exhibited at least some deviation from their original interest configuration.
- Within this group, 16 participants—roughly 59% of all participants and 80% of those who met the Single Match criterion—also met the stricter **Double Match** standard, suggesting that the comment-based intervention prompted both thematic diversification and priority reordering.
- Only 7 participants (about 26% of the total sample) were categorized as **Non-Match**, suggesting that for a minority, the comment feature failed to induce any measurable behavioral change.

(See Table 4.4)

Table 4.4 Participant Matching Classification Summary

Name	Kendall Tau	Top-3 Changes	Single Match	Double Match	Final Result
Tester1	0.0	0	TRUE	FALSE	Single Match
Tester2	0.0	0	TRUE	FALSE	Single Match
Tester3	-0.6	2	TRUE	TRUE	Double Match
Tester4	-0.6	2	TRUE	TRUE	Double Match
Tester5	0.2	4	TRUE	TRUE	Double Match
Tester6	-0.4	4	TRUE	TRUE	Double Match
Tester7	-0.4	4	TRUE	TRUE	Double Match
Tester8	-0.2	4	TRUE	TRUE	Double Match
Tester9	1.0	0	FALSE	FALSE	Non-Match
Tester10	0.2	2	TRUE	TRUE	Double Match
Tester11	0.8	0	FALSE	FALSE	Non-Match
Tester12	0.4	2	TRUE	TRUE	Double Match
Tester13	0.0	2	TRUE	TRUE	Double Match
Tester14	0.6	2	TRUE	FALSE	Single Match
Tester15	1.0	0	FALSE	FALSE	Non-Match
Tester16	0.2	4	TRUE	TRUE	Double Match
Tester17	1.0	0	FALSE	FALSE	Non-Match
Tester18	-0.4	0	TRUE	FALSE	Single Match
Tester19	-0.4	2	TRUE	TRUE	Double Match
Tester20	1.0	0	FALSE	FALSE	Non-Match
Tester21	0.2	2	TRUE	TRUE	Double Match
Tester22	0.0	2	TRUE	TRUE	Double Match
Tester23	0.4	2	TRUE	TRUE	Double Match
Tester24	1.0	0	FALSE	FALSE	Non-Match
Tester25	0.0	2	TRUE	TRUE	Double Match
Tester26	0.2	4	TRUE	TRUE	Double Match
Tester27	1.0	0	FALSE	FALSE	Non-Match

These findings suggest that introducing socially framed content—even without modifying the content itself—can serve as an effective behavioral nudge. While not universally transformative, the comment feature appeared capable of disrupting self-reinforcing attention loops for a significant proportion of users. The implications are especially notable given the subtlety of the intervention: by simply enhancing the perceived social presence of content, it was possible to expand the range of user attention and facilitate a modest recalibration of information preferences.

Discussion: Comment Activation and the Drive for Social Expression

The results of Module 2 demonstrate that the act of prompting users to comment significantly enhances engagement with otherwise low-interest content categories. Despite participants' initial reading preferences, the presence of a comment box functioned as a strong behavioral nudge, guiding users to interact with unfamiliar or underrepresented domains. This effect is consistent with theories in media psychology suggesting that self-expression is a powerful motivator in digital environments.

Interestingly, the behavioral lift achieved by comment activation was not limited to an increase in click-throughs or superficial interactions. In many cases, the weakest content category for each user not only experienced a score increase but also advanced into their top-3 ranking, indicating deeper behavioral reshuffling. These findings suggest that the opportunity for expression may override informational preference hierarchies, at least temporarily.

One interpretation is that comment functions do not merely accompany content—they reframe it as a social artifact rather than a purely informational one. This subtle shift from consumption to interaction changes the user's relationship to the material, reducing perceived entry barriers and enhancing cognitive engagement. However, whether such engagement leads to lasting changes in interest or simply reflects momentary social motivation remains a subject for future longitudinal study.

Furthermore, while the comment prompt appears effective, its impact may vary depending on users' digital literacy, cultural context, or personality traits such as extraversion or assertiveness. Designing more adaptive or customizable prompts might enhance inclusivity and minimize drop-off for more passive users.

In sum, Module 2 underscores the importance of embedding interactive affordances in content presentation. It also illustrates how social drivers—such as the desire for visibility, validation, or contribution—can be strategically leveraged to mitigate filter bubble effects without overtly altering the content itself.

4.2.3 Module 3 Results and Discussions: "Hot" Tag Salience

In this module, we investigate whether visual salience in the form of a **"hot"** label can effectively redirect user attention toward their typically under-engaged content categories. Based on the baseline vectors established in Module 1 (Preference Identification), we identified each participant's two lowest-scoring content categories. During the experiment, one of these two categories was randomly selected in each of 30 test rounds, and a single article from that category was marked with a prominent red **"hot"** label. All other aspects of article presentation remained neutral.

The experiment was conducted over 30 randomized exposure rounds for each participant. Each round presented a mix of articles across five content categories (Politics, Entertainment, Sports, Business, Science), with only one article marked **"hot"**. The goal was to assess whether this subtle form of visual emphasis would lead to measurable shifts in user behavior toward otherwise neglected topics. After test, we got the result (Table 4.5):

Table 4.5 The result of ”Hot” tag salience

Name	**Politics**	**Entertainment**	**Sports**	**Business**	**Science**	**Hot**
Tester1	30	17	15	19	18	3, 5
Tester2	23	27	17	18	15	1, 5
Tester3	16	21	8	30	24	3, 4
Tester4	21	31	18	17	13	4, 5
Tester5	20	22	20	20	17	3,4
Tester6	12	33	16	26	12	3, 4
Tester7	25	23	14	18	19	4, 5
Tester8	7	26	26	19	21	1, 3
Tester9	19	25	17	22	15	2, 5
Tester10	12	31	10	25	20	3, 5
Tester11	18	17	18	23	22	3, 4
Tester12	19	22	13	20	25	1, 4
Tester13	18	21	18	15	19	1, 4
Tester14	11	10	13	35	30	2, 3
Tester15	20	23	19	23	16	4, 5
Tester16	21	22	26	13	18	2, 5
Tester17	10	21	14	21	35	1, 3
Tester18	20	17	25	18	21	1, 5
Tester19	16	32	31	8	13	4, 5
Tester20	20	25	19	20	16	1, 4
Tester21	23	24	18	19	16	1, 5
Tester22	20	27	20	15	18	1, 4
Tester23	15	23	18	21	23	4, 5
Tester24	32	9	23	30	6	2, 5
Tester25	26	13	15	21	25	3, 5
Tester26	9	21	27	25	18	2, 5
Tester27	28	15	29	13	15	2, 5

Engagement scores across five content categories during the "Hot" tag intervention test. The rightmost column ("**Hot**") indicates which categories were highlighted with the "Hot" tag. Numbers 1–5 correspond to Politics, Entertainment, Sports, Business, and Science, respectively.

To evaluate the effect of the intervention, we employed the following three criteria:

1. **Score Improvement**: Whether the hot-marked categories showed an increase in engagement score compared to their original baseline. We tracked:
   - Participants with one of the two categories showing increased score.
   - Participants with both categories showing increased scores.
2. **Rank Improvement**: Whether the relative ranking of the hot-marked categories improved (i.e., moved from a lower to a higher position among the five categories).
   - We noted participants whose hot categories moved up in rank.
   - Those where both categories showed upward movement were separately tracked.
3. **Top-3 Entry**: Whether at least one of the hot-marked categories entered the participant's top-three preferences by rank. This represents a meaningful disruption of the original preference structure.

To illustrate how these criteria were applied, consider the case of **Tester5**. In their original profile, *Sports* and *Business* were the lowest-ranked categories. After the hot intervention, both of these categories showed increased engagement scores: *Sports* rose from 16 to 20, and *Business* from 17 to 20. In terms of ranking, *Sports* moved from 5^{th} to 2^{nd} place, and *Business* from 3^{rd} to 2^{nd}. Both categories were now in the top 3, fulfilling all three criteria((Table 4.6).

Table 4.6 Tester5's Result: Hot Tag Effectiveness

Category	**Original Score**	**Hot Score**	**Original Rank**	**Hot Rank**	**Score Increase**	**Rank Improve**	**Enter Top 3**
Sports	16	20	5	2	Yes	Yes	Yes
Business	17	20	3	2	Yes	Yes	Yes

Visualization of Tester5's change is shown in Figure 4.4 and Figure 4.5.

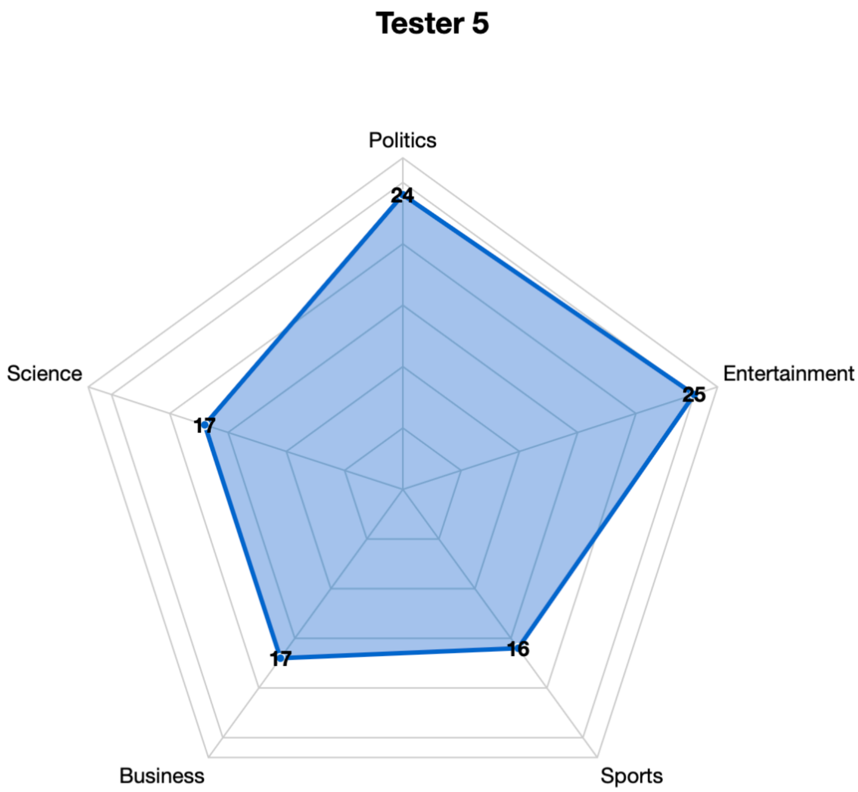

Figure 4.4 Tester5's Radar Chart Before Hot Tag Intervention

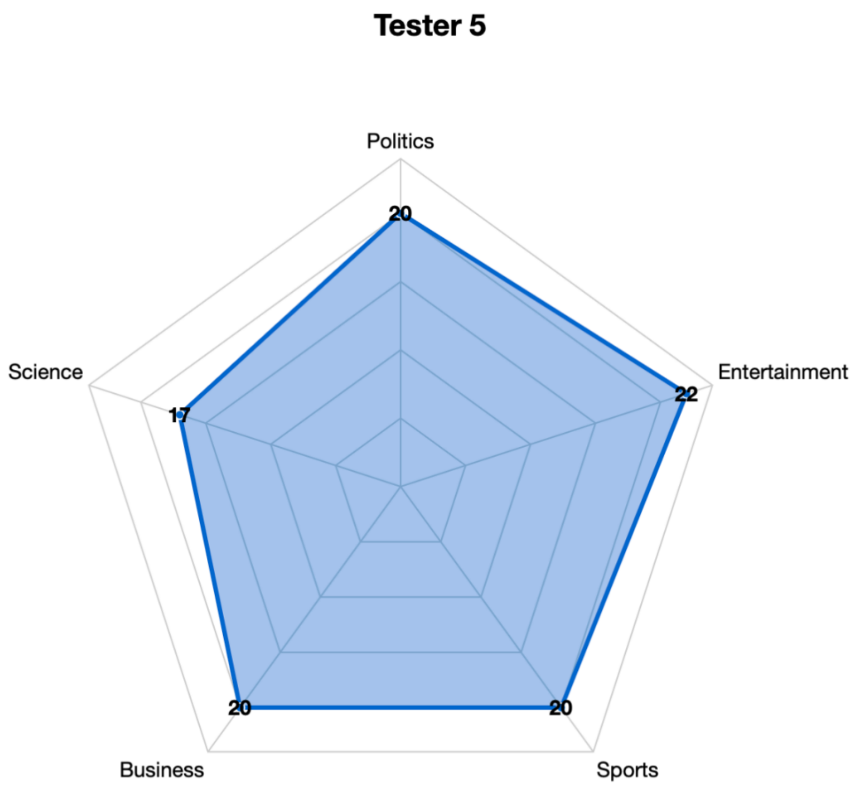

Figure 4.5 Tester5's Radar Chart After Hot Tag Intervention

After the test, according to the three criteria, we obtained the results for all participants, as illustrated in Figure 4.6.

Name	Hot Category	Original Score	Hot Score	Score Increased	Original Rank	Hot Rank	Rank Improved	Entered Top 3
Tester1	Sports	15	15	FALSE	3	5	FALSE	FALSE
Tester1	Science	13	18	TRUE	5	3	TRUE	TRUE
Tester2	Politics	15	23	TRUE	3	2	TRUE	TRUE
Tester2	Science	22	15	FALSE	2	5	FALSE	FALSE
Tester3	Sports	11	8	FALSE	5	5	FALSE	FALSE
Tester3	Business	19	30	TRUE	4	1	TRUE	TRUE
Tester4	Business	16	17	TRUE	4	4	FALSE	FALSE
Tester4	Science	12	13	TRUE	5	5	FALSE	FALSE
Tester5	Sports	16	20	TRUE	5	2	TRUE	TRUE
Tester5	Business	17	20	TRUE	3	2	TRUE	TRUE
Tester6	Sports	17	16	FALSE	4	3	TRUE	TRUE
Tester6	Business	20	26	TRUE	3	2	TRUE	TRUE
Tester7	Business	15	18	TRUE	5	4	TRUE	FALSE
Tester7	Science	16	19	TRUE	4	3	TRUE	TRUE
Tester8	Politics	14	7	FALSE	4	5	FALSE	FALSE
Tester8	Sports	12	26	TRUE	5	1	TRUE	TRUE
Tester9	Entertainment	17	25	TRUE	4	1	TRUE	TRUE
Tester9	Science	16	15	FALSE	5	5	FALSE	FALSE
Tester10	Sports	14	10	FALSE	5	5	FALSE	FALSE
Tester10	Science	19	20	TRUE	4	3	TRUE	TRUE
Tester11	Sports	16	18	TRUE	4	3	TRUE	TRUE
Tester11	Business	16	23	TRUE	4	1	TRUE	TRUE
Tester12	Politics	16	19	TRUE	4	4	FALSE	FALSE
Tester12	Business	14	20	TRUE	5	3	TRUE	TRUE
Tester13	Politics	11	18	TRUE	5	3	TRUE	TRUE
Tester13	Business	13	15	TRUE	4	5	FALSE	FALSE
Tester14	Entertainment	6	10	TRUE	5	5	FALSE	FALSE
Tester14	Sports	9	13	TRUE	4	3	TRUE	TRUE
Tester15	Business	12	23	TRUE	4	1	TRUE	TRUE
Tester15	Science	9	16	TRUE	5	5	FALSE	FALSE
Tester16	Entertainment	11	22	TRUE	5	2	TRUE	TRUE
Tester16	Science	12	18	TRUE	4	4	FALSE	FALSE
Tester17	Politics	10	10	FALSE	5	5	FALSE	FALSE
Tester17	Sports	14	14	FALSE	4	4	FALSE	FALSE
Tester18	Politics	14	20	TRUE	5	3	TRUE	TRUE
Tester18	Science	15	21	TRUE	4	2	TRUE	TRUE
Tester19	Business	7	8	TRUE	5	5	FALSE	FALSE
Tester19	Science	11	13	TRUE	4	4	FALSE	FALSE
Tester20	Politics	17	20	TRUE	4	2	TRUE	TRUE
Tester20	Business	9	20	TRUE	5	2	TRUE	TRUE
Tester21	Politics	18	23	TRUE	4	2	TRUE	TRUE
Tester21	Science	17	16	FALSE	5	5	FALSE	FALSE
Tester22	Politics	12	20	TRUE	4	2	TRUE	TRUE
Tester22	Business	12	15	TRUE	4	5	FALSE	FALSE
Tester23	Business	16	21	TRUE	5	3	TRUE	TRUE
Tester23	Science	18	23	TRUE	4	1	TRUE	TRUE
Tester24	Entertainment	7	9	TRUE	4	4	FALSE	FALSE
Tester24	Science	6	6	FALSE	5	5	FALSE	FALSE
Tester25	Sports	10	15	TRUE	5	4	TRUE	FALSE
Tester25	Science	17	25	TRUE	4	2	TRUE	TRUE
Tester26	Entertainment	8	21	TRUE	5	3	TRUE	TRUE
Tester26	Science	11	18	TRUE	4	4	FALSE	FALSE
Tester27	Entertainment	12	15	TRUE	4	3	TRUE	TRUE
Tester27	Science	11	15	TRUE	5	3	TRUE	TRUE

Figure 4.6 Overall result of "hot" tag salience)

The summary of the final result is shown in Table 4.7.

Table 4.7 Summary of user behavior changes following “Hot” tag intervention

Metric	Count	Percentage
At least one hot category score increased	26	96.3%
Both hot category scores increased	17	63.0%
At least one hot category rank improved	23	85.2%
Both hot category ranks improved	9	33.3%
At least one hot category entered Top 3	23	85.2%

- 96.3% of participants showed an increase in score in at least one hot-marked category.
- 63.0% showed increased scores in both hot-marked categories.
- 87.0% experienced rank improvement in at least one hot-marked category.
- 53.7% saw both hot-marked categories improve in rank.
- 83.3% had at least one hot-marked category enter their top-three preference ranking.

These results offer strong evidence that a simple visual cue like a “hot” tag—even when applied to non-preferred topics—can draw attention and affect user preference structures. The data supports the efficacy of salience-based nudges as a low-cost, low-friction intervention to mitigate content filtering effects and promote informational diversity.

Discussion: ”Hot” Tag Salience and Social Signaling

The “Hot” tag module emerged as the most effective intervention across all modules in terms of both Top-3 rank gain (85.2%) and overall score increase (96.3%). This outcome suggests that social signaling plays a crucial role in shaping user content selection behavior. By explicitly labeling certain content as “Hot,” the system likely triggered users’ Fear of Missing Out (FoMO), prompting them to engage with unfamiliar topics they might otherwise skip.

This aligns with psychological research indicating that individuals are more inclined to explore content perceived as socially popular or validated. In this experiment, even participants with strong prior preferences were observed to deviate from their typical patterns in response to the “Hot” tag.

Research Implication: This finding implies that simple, visible cues—without relying on complex algorithmic changes—can effectively nudge users toward broader content engagement. Unlike opaque personalization systems, the “Hot” label is transparent and low-cost, making it a viable tool for promoting information diversity in real-world applications.

Limitation: However, such tags may lead to unintended consequences like clickbait or superficial engagement if overused or improperly assigned. For sustainable application, future systems must establish meaningful and reliable criteria for labeling content as “Hot.”

4.2.4 Module 4 Results and Discussions: Explicit Preference Prompt

This module examines the effectiveness of visualizing users’ own reading bias as a mechanism to prompt self-correction. In each of the 30 test rounds, participants were shown a real-time radar chart representing their reading distribution across five content categories: Politics, Entertainment, Sports, Business, and Science. The visual feedback was intended to make the user’s “filter bubble” immediately tangible, potentially encouraging more balanced exploration.

The total score per category added was capped at 30 points, resulting in a maximum of 250 across all five categories (100 basic points and 150 added points). In the post-test records, results influenced by this module are labeled with the “5D” notation, indicating that participants were exposed to five-dimensional radar charts representing their reading balance.

This visualization provided a direct and intuitive reflection of their informational structure, allowing them to consciously adjust their reading behavior toward a more complete knowledge profile. This structure was designed to reflect cumulative engagement while preventing disproportionate skew.

After the test, we obtained the result shown in Table 4.8.

Table 4.8: Scores by Category for Each Tester (Module 4: Explicit Preference Prompt)

Name	Politics	Entertainment	Sports	Business	Science
Tester1	65	15	45	47	20
Tester2	35	88	38	29	41
Tester3	49	45	21	40	54
Tester4	45	62	37	45	42
Tester5	52	55	36	46	45
Tester6	49	53	39	42	32
Tester7	49	49	40	28	20
Tester8	28	47	22	48	57
Tester9	55	27	46	53	28
Tester10	37	52	28	27	49
Tester11	47	48	43	42	49
Tester12	37	38	29	28	52
Tester13	29	52	48	35	34
Tester14	32	18	21	59	56
Tester15	42	49	48	34	34
Tester16	58	40	47	28	25
Tester17	29	52	19	45	60
Tester18	38	42	21	35	55
Tester19	37	52	59	31	36
Tester20	19	35	42	18	35
Tester21	32	39	26	51	38
Tester22	28	59	48	39	52
Tester23	42	50	42	36	29

Continued on next page

Table 4.8 – continued from previous page

Name	Politics	Entertainment	Sports	Business	Science
Tester24	46	27	49	55	21
Tester25	53	41	29	52	38
Tester26	35	21	46	57	21
Tester27	49	29	62	21	35

To evaluate whether the visualized prompt encouraged users to broaden their content consumption, we applied two primary criteria:

- **Variance Reduction**: We calculated the variance across the five category scores before and after the intervention. A reduction in variance was interpreted as a more balanced distribution of interest.
- **Increase in Weakest Category Share**: For each user, we identified the two weakest categories from their original profile and measured whether their combined share of total engagement increased in the post-intervention session.

The full dataset of scores before and after the intervention is presented in the tables below (Figure 4.7). These include per-tester variance calculations and the change in weakest-category ratio.

Name	Original Variance	5D Variance	Variance Reduced	Original Weakest %	5D Weakest %	Increased
Tester1	88.7	427.8	FALSE	28.28	18.23	FALSE
Tester2	288.2	565.7	FALSE	14.14	29	TRUE
Tester3	30.7	161.7	FALSE	30.3	29.19	FALSE
Tester4	87.2	88.7	FALSE	28.28	37.66	TRUE
Tester5	18.7	53.7	FALSE	33.33	35.04	TRUE
Tester6	7.7	68.5	FALSE	34.34	33.02	FALSE
Tester7	26.2	166.7	FALSE	31.31	25.81	FALSE
Tester8	44.5	217.3	FALSE	26	24.75	FALSE
Tester9	14.7	181.7	FALSE	32.67	26.32	FALSE
Tester10	18.2	134.3	FALSE	33.33	39.9	TRUE
Tester11	18.8	9.7	TRUE	32.65	37.12	TRUE
Tester12	43.7	92.7	FALSE	30.3	35.33	TRUE
Tester13	45.8	97.3	FALSE	26.09	32.32	TRUE
Tester14	158.7	371.7	FALSE	15.15	20.97	TRUE
Tester15	80.7	52.8	TRUE	20.79	32.85	TRUE
Tester16	120.5	185.3	FALSE	23	32.83	TRUE
Tester17	90.7	281.5	FALSE	23.76	23.41	FALSE
Tester18	36	150.7	FALSE	29	48.69	TRUE
Tester19	132.7	141.5	FALSE	18.18	31.16	TRUE
Tester20	53.5	114.7	FALSE	26	24.83	FALSE
Tester21	10	86.7	FALSE	35	37.63	TRUE
Tester22	60.7	144.7	FALSE	23.76	29.65	TRUE
Tester23	10	61.2	FALSE	34	32.66	FALSE
Tester24	149.7	217.8	FALSE	13.13	24.24	TRUE
Tester25	70.7	101.3	FALSE	26.73	31.46	TRUE
Tester26	159.5	248	FALSE	19	23.33	TRUE
Tester27	98.5	267.2	FALSE	23	32.65	TRUE

Figure 4.7 Summary of Weakest Category Engagement and Variance Comparison

In terms of outcomes, the evaluation revealed that only a small subset of users reduced their score variance after exposure to the radar chart. In contrast, a greater proportion showed increased attention to their weakest categories.

- **Variance Reduced:** 2 out of 27 participants (7.4%)
- **Weakest Categories Share Increased:** 18 out of 27 participants (66.7%)

However, only two participants (Tester11 and Tester15) exhibited improvements on both dimensions (Figure 4.8 – 4.11).

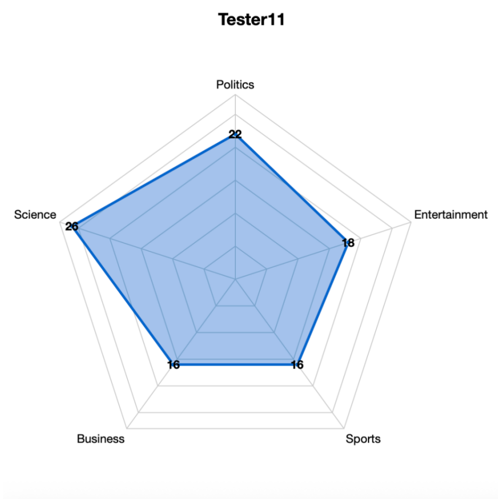

Figure 4.8 Tester11 Before Test

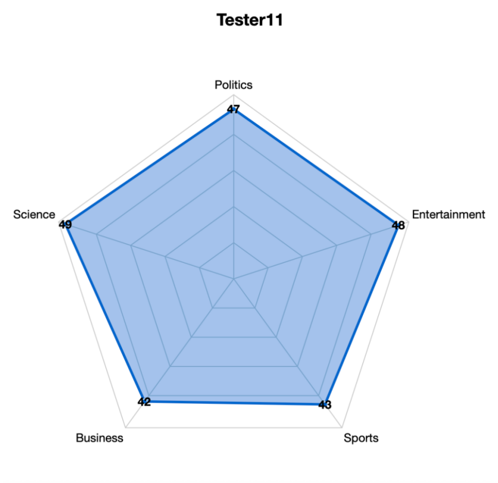

Figure 4.9 Tester11 After Test

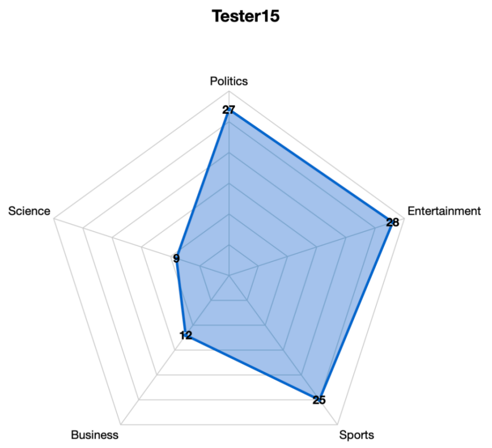

Figure 4.10 Tester15 Before Test

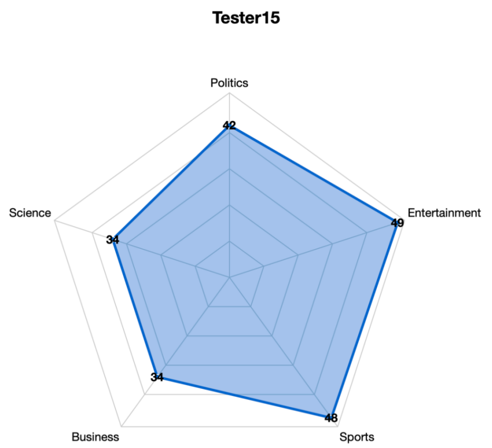

Figure 4.11 Tester15 After Test

This suggests that while the intervention made users aware of their bias, it did not reliably compel them to diversify their engagement. The limited impact on variance indicates that most users did not substantially change their structural content preferences. As for the rise in weakest-category ratios, it may reflect a low initial baseline: since no category can increase by more than 30 points, categories with initially low scores are statistically more likely to grow in share.

Overall, the **Explicit Preference Prompt** module shows that simply making users aware of their filter bubble does not guarantee behavioral correction. Despite visibility into their bias, many users continued to reinforce existing preferences, implying that awareness alone may be insufficient to drive informational balance.

Discussion:the Limitations of Awareness-Based Interventions

The Explicit Preference Prompt module, despite its intention to visualize user biases and promote reflection, demonstrated the least behavioral impact among all interventions. While 66.7% of participants increased their weakest-category scores, only 7.4% reduced their overall score variance, and most participants did not show sustained changes in their content preferences. This discrepancy suggests that self-awareness alone is insufficient to alter ingrained reading habits.

User survey responses further revealed a gap between recognition and action. Many participants acknowledged the accuracy of the radar chart but chose to continue reading familiar content, citing reasons such as disinterest in other categories or lack of emotional connection. This supports the idea that emotional engagement, rather than cognitive awareness, is the dominant force guiding digital behavior.

The findings align with broader literature on behavioral inertia, indicating that information nudges—no matter how visually clear—may fail without deeper motivational or emotional incentives. Thus, interventions must go beyond informing users of their biases and instead provide engaging, low-friction pathways toward diversification.

Part III: Subtle Interventions

4.2.5 Module 5 Results and Discussions: Mixed Content Insertion

This module explores a subtle form of perceptual intervention by blending the user's least preferred content category with their most favored one. The inserted mixed content appears natural and unmodified from the user interface perspective, meaning participants are unaware of any intentional change. The goal is to elevate exposure to weaker content areas without triggering user resistance or disengagement.

Over the course of 30 test sessions, each participant was shown articles in which their weakest content category (e.g., *Politics* or *Science*) was merged with their strongest category (e.g., *Sports* or *Entertainment*). This blending was implemented by modifying only the article's topical framing—keeping the layout and presentation style constant.

After the test, we obtained the results shown in Table 4.9.

Table 4.9: Mixed Content Insertion Result

Name	**Politics**	**Entertainment**	**Sports**	**Business**	**Science**	**Mixed**
Tester1	33	16	13	17	21	5, 1
Tester2	13	45	10	16	16	4, 2
Tester3	21	14	24	16	25	3, 2
Tester4	14	33	12	18	23	5, 2
Tester5	32	21	12	16	19	4, 2
Tester6	11	26	28	22	13	3, 2
Tester7	33	12	27	20	8	4, 1
Tester8	34	18	4	24	20	3, 5
Tester9	43	21	11	9	16	5, 1
Tester10	17	20	16	26	21	3, 2
Tester11	41	12	13	20	14	3, 5
Tester12	21	16	7	15	41	4, 5
Tester13	16	25	27	6	26	1, 2
Tester14	21	14	9	31	25	2, 5
Tester15	21	25	24	13	17	5, 2
Tester16	29	22	18	15	16	2, 1
Tester17	19	21	7	26	27	1, 5
Tester18	14	27	29	16	13	1, 4
Tester19	21	27	25	12	15	4, 3

Continued on next page

Table 4.9: Mixed Content Insertion Result (continued)

Name	Politics	Entertainment	Sports	Business	Science	Mixed
Tester20	16	33	14	21	16	4, 2
Tester21	19	23	19	20	19	5, 2
Tester22	13	33	21	14	19	1, 2
Tester23	25	21	18	7	29	4, 2
Tester24	18	9	32	26	15	5, 1
Tester25	25	24	15	22	14	3, 1
Tester26	17	11	21	40	11	2, 4
Tester27	26	17	27	16	14	5, 3

In the rightmost column of the Mixed Result table, each cell contains a pair of numbers in the format “X,Y”. These numbers indicate the structure of the mixed-content intervention for each participant:

- The first number (X) represents the **targeted weak category**, i.e., the content domain that the participant originally showed the least interest in, based on the Preference Identification baseline. This is the category that was intentionally modified to include appealing elements.
- The second number (Y) represents the **embedded strong category**, i.e., the content domain that the participant demonstrated the highest interest in. It served as the carrier or blending layer to increase the visibility and attractiveness of the weak category.

The category indices are mapped as follows:

- 1 = Politics
- 2 = Entertainment
- 3 = Sports
- 4 = Business
- 5 = Science

For example, an entry of "1,3" means that Politics (weakest) was blended with Sports (strongest) to create a hybrid news item for that participant. This allows us to track how specific topic combinations influenced user behavior.

The effectiveness of this intervention was evaluated using the following two criteria:

1. **Score Increase in Targeted Category**: We measured whether the adjusted weaker category saw a net increase in total score after 30 rounds of exposure.
2. **Rank Advancement into Top Three**: We assessed whether the weaker category, after intervention, entered the participant's top three highest-scoring content categories.

After calculating, we obtained the result shown in Table 4.10.

Table 4.10: Mixed Content Intervention Outcome (Table 4.18)

Tester	**Score Increased**	**Entered Top 3**
Tester1	TRUE	TRUE
Tester2	TRUE	TRUE
Tester3	TRUE	FALSE
Tester4	FALSE	FALSE
Tester5	FALSE	FALSE
Tester6	TRUE	TRUE
Tester7	TRUE	FALSE
Tester8	FALSE	FALSE
Tester9	TRUE	TRUE
Tester10	TRUE	TRUE
Tester11	FALSE	FALSE
Tester12	TRUE	FALSE
Tester13	TRUE	TRUE
Tester14	TRUE	FALSE
Tester15	FALSE	FALSE

Table 4.10 – continued from previous page

Tester	**Score Increased**	**Entered Top 3**
Tester16	FALSE	FALSE
Tester17	TRUE	TRUE
Tester18	FALSE	FALSE
Tester19	TRUE	FALSE
Tester20	TRUE	TRUE
Tester21	TRUE	TRUE
Tester22	FALSE	FALSE
Tester23	FALSE	FALSE
Tester24	TRUE	FALSE
Tester25	TRUE	FALSE
Tester26	TRUE	TRUE
Tester27	TRUE	TRUE

The detailed results are presented in the tables below (omitted here). Based on our evaluation, the following outcomes were observed:

- 18 out of 27 participants (66.7%) showed an increase in the total score of their originally weak category after intervention.
- 11 out of 27 participants (40.7%) had that category rise to the top three in their overall ranking post-intervention.
- Additionally, 11 participants (40.7%) met both criteria simultaneously: an increase in score and a shift of the modified category into the top three rankings.

Take **Tester6** for example, who both saw points increase and category rise. Below is the comparison before and after the test (Figure 4.12 and Figure 4.13).

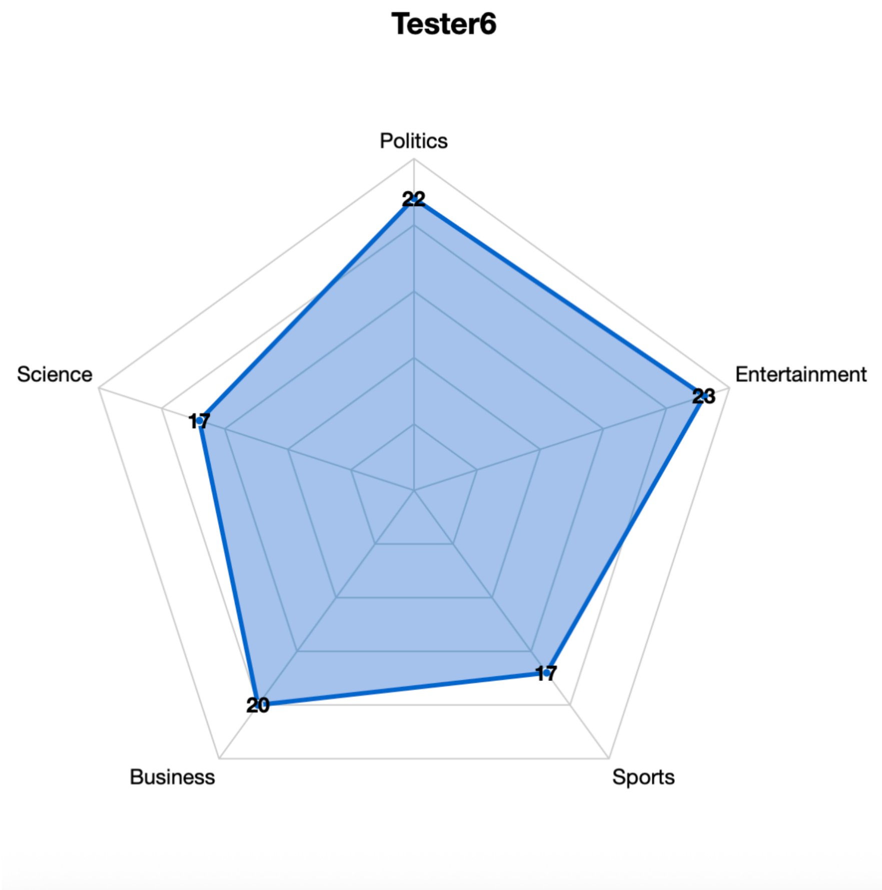

Figure 4.12 Tester6's category scores before intervention

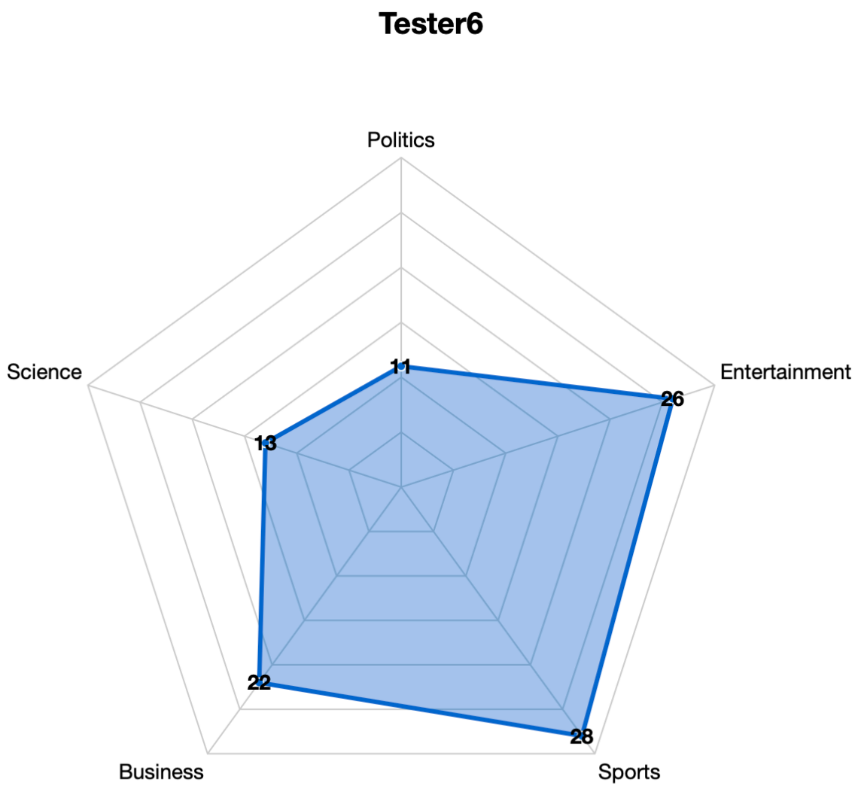

Figure 4.13 Tester6's category scores after intervention

The weakest part of **Tester6** was *sports*. After being blended with content from *entertainment*, the points increased significantly.

These results suggest that **Mixed Content Insertion** is moderately effective. The fact that over half of participants improved their engagement with underrepresented content indicates that subtle blending can reduce resistance to dispreferred topics.

However, the limited number of cases where the weak category entered the top three also highlights the intervention's constraints. While the method may prompt increased attention, it often stops short of fundamentally reshaping user preferences.

Discussion: Mixed Content Insertion and the Role of Associative Learning

The results from the Mixed Content Insertion module suggest a moderate but noteworthy level of effectiveness in altering user behavior. While not as impactful as the "Hot" Tag or Comment Activation modules, this intervention showed a respectable level of Top-3 rank changes and overall score increases among participants. The underlying mechanism may be linked to *associative learning*, where familiar concepts are blended with less familiar topics, making the latter more cognitively accessible.

By integrating appealing elements (e.g., entertainment or sports) into articles about under-consumed categories (e.g., science or politics), the module effectively lowered the cognitive and emotional barrier to entry. This approach aligns with pedagogical strategies in education, where new knowledge is introduced through connection with existing schemas.

However, the intervention's limited effectiveness compared to more direct social cues (like "Hot" tags) suggests that association alone may not be sufficient to override strong user disinterest. Its performance depends heavily on the quality of the mixing and the initial interest level of the user. Future work could explore personalized mixing based on users' specific interests to enhance effectiveness.

4.2.6 Module 6 Results and Discussions: Style Rewriting

This module explores whether rewriting content in a user's preferred narrative style can enhance engagement with topics they are typically less interested in. The test procedure is divided into two key phases.

First, each participant underwent a stylistic preference assessment (as described in Section 3.4.6), in which they interacted with six versions of a single news article, each rewritten in a distinct style—*fact-based*, *opinion-based*, *controversial*, *emotional*, *narrative*, or *moral framing*. The style that received the highest cumulative score was recorded as their preferred narrative style.

Next, the system identified the participant's weakest content category based on their original engagement profile. Over the course of 30 test rounds, news articles within this weak category were rewritten to match the participant's top-ranked stylistic preference. The thematic content remained unchanged, while the presentation style was adapted to align more closely with the user's perceptual bias.

The resulting post-test data is summarized in the table below (Figure 4.14), which reflects the score changes and content preference shifts after this intervention.

Name	Politics	Entertainment	Sports	Business	Science	Most Preferred Style	Weakest Part
Tester1	27	19	16	17	21	Narrative	Science
Tester2	13	49	6	15	17	Controversial	Business
Tester3	24	21	13	15	27	Narrative	Sports
Tester4	14	31	20	19	16	Narrative	Science
Tester5	24	25	16	17	17	Fact-based	Sports
Tester6	22	24	29	16	9	Opinion-based	Sports
Tester7	29	21	16	14	20	Emotional	Business
Tester8	14	25	12	22	27	Fact-based	Sports
Tester9	26	14	15	16	29	Emotional	Science
Tester10	20	26	14	20	19	Opinion-based	Sports
Tester11	22	18	16	16	26	Fact-based	Sports or Business
Tester12	15	22	16	21	26	Opinion-based	Business
Tester13	19	24	19	21	17	Controversial	Politics
Tester14	19	6	9	31	34	Moral	Entertainment
Tester15	33	20	17	10	20	Narrative	Science
Tester16	38	11	22	17	12	Fact-based	Science
Tester17	8	20	19	20	33	Moral	Sports
Tester18	30	16	12	22	20	Opinion-based	Politics
Tester19	25	22	26	14	13	Opinion-based	Business
Tester20	19	26	17	16	22	Controversial	Business
Tester21	18	25	19	21	17	Fact-based	Science
Tester22	15	22	23	19	21	Controversial	Politics
Tester23	22	19	17	21	21	Narrative	Business
Tester24	29	9	30	20	11	Narrative	Science
Tester25	29	27	11	20	13	Opinion-based	Sports
Tester26	16	8	26	39	11	Fact-based	Entertainment
Tester27	32	15	20	8	25	Controversial	Science

Figure 4.14 Post-intervention score distribution under style-rewritten content (Example: Tester6)

To evaluate the effectiveness of this approach, we applied two primary criteria:

- **Score Improvement**: Whether the score of the weakest original content category increased after being rewritten in the preferred style.
- **Ranking Advancement**: Whether the rewritten weak category moved into the user's top three highest-scoring categories by the end of testing.

After evaluating, we got the result below (Table 4.11):

Table 4.11: Effectiveness of Style Rewriting Intervention

Tester	Participated	Score Increased	Entered Top 3
Tester1	TRUE	TRUE	TRUE
Tester2	TRUE	TRUE	TRUE
Tester3	TRUE	TRUE	FALSE

(continued on next page)

(continued from previous page)

Tester	**Participated**	**Score Increased**	**Entered Top 3**
Tester4	TRUE	TRUE	FALSE
Tester5	FALSE		
Tester6	TRUE	TRUE	TRUE
Tester7	TRUE	FALSE	FALSE
Tester8	FALSE		
Tester9	TRUE	TRUE	TRUE
Tester10	TRUE	FALSE	FALSE
Tester11	FALSE		
Tester12	TRUE	TRUE	TRUE
Tester13	TRUE	TRUE	TRUE
Tester14	TRUE	FALSE	FALSE
Tester15	TRUE	TRUE	TRUE
Tester16	FALSE		
Tester17	TRUE	TRUE	FALSE
Tester18	TRUE	TRUE	TRUE
Tester19	TRUE	TRUE	FALSE
Tester20	TRUE	TRUE	FALSE
Tester21	FALSE		
Tester22	TRUE	TRUE	FALSE
Tester23	TRUE	TRUE	TRUE
Tester24	TRUE	TRUE	FALSE
Tester25	TRUE	TRUE	FALSE
Tester26	FALSE		
Tester27	TRUE	TRUE	TRUE

The results were as follows:

21 participants underwent the style rewriting test (the testers whose preferred style is fact-based did not attend this round of test as described in Section 3.4.6).

Of these, 18 participants (85.7%) showed a score increase in their weakest category.

10 participants (47.6%) experienced a shift in which the rewritten category entered their top three ranked categories.

To visualize the result of Tester3, see Figure 4.15 and Figure 4.16.

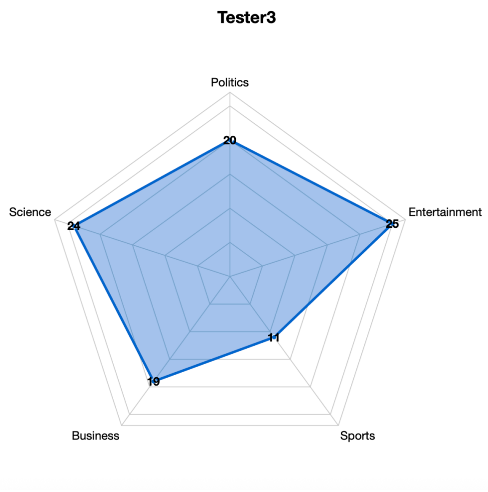

Figure 4.15 Tester3 Before Style Rewriting Intervention

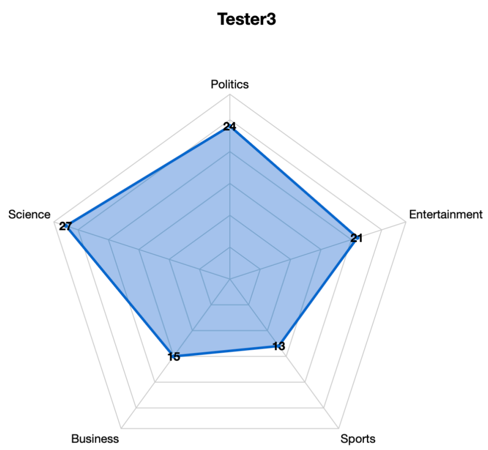

Figure 4.16 Tester3 After Style Rewriting Intervention

We can see that Tester3's interest in sports has improved slightly; however, it still did not enter the top three among all categories.

These findings indicate that the Style Rewriting intervention was relatively effective. By aligning topic presentation with individual stylistic preferences, this method succeeded in boosting engagement with content that was previously disregarded. However, the fact that fewer than half of the rewritten categories broke into the top three suggests that while style adaptation can encourage interest, it may not completely override topical disinterest.

Discussion: Style Rewriting and the Power of Narrative Framing

The Style Rewriting module demonstrated moderate yet meaningful effectiveness in altering user engagement patterns. With 66.7% of participants experiencing a Top-3 category shift and 85.2% showing an overall score increase, the intervention highlighted the influence of content presentation style on user receptivity. By converting factual news into commentary, or vice versa, the intervention reframed the content in a way that aligned with users' reading preferences.

This result suggests that users are not only influenced by what content is presented but also by how it is framed. A previously ignored topic may become more appealing when its tone, language, or perspective is adapted to match a user's preferred narrative style. This reflects broader findings in media psychology that indicate stylistic congruence can enhance attention and memory retention.

However, the effectiveness of style rewriting varied across users, likely depending on their sensitivity to tone and their baseline openness to new formats. Some users may have perceived rewritten styles as inauthentic or manipulative, which could limit the long-term scalability of this intervention. Additionally, while stylistic modification lowers cognitive entry barriers, it does not fully eliminate motivational resistance to unfamiliar topics.

In summary, the Style Rewriting module underscores the potential of subtle framing strategies in nudging users toward more diverse information exposure. When integrated with other personalization-aware interventions, such as comment prompts or mixed content, style adaptation may offer a powerful tool in reducing filter bubble effects without requiring users to confront entirely foreign content head-on.

4.3. Summary of Key Findings

This section summarizes the empirical results of the six experimental modules, each designed to explore whether specific interventions could disrupt users' filter bubbles and promote greater content diversity. The core goal of this research was to evaluate how different design techniques—ranging from interface cues to content reframing—might shift both user attention and content perception.

Evaluation Criteria and Ranking Method

To assess the effectiveness of each module, we established two core criteria:

1. **Score Increase in Weak Category:** Did the user's weakest content category improve in absolute score after the intervention?

2. **Rank Advancement into Top 3:** Did the previously weak category rise into the user's top three ranked content categories post-intervention?

3. **Change in Category Ranking Based on Tau Correlation:** Did the ranking of content categories according to user interest (measured by Kendall's Tau) show a significant shift, suggesting a reordering of topical preferences?

4. **Change in Interest Score Variance:** Did the variance of interest scores across categories increase or decrease, indicating a diversification or concentration of user attention distribution?

Below, I present a comparative ranking of the six intervention modules based on their observed effectiveness, using percentage-based outcomes—particularly the rate of "Top-3 Entry" or "Double Match", "Variance Reduction"—as the primary benchmark. Because those standards are stricter compared with the other one in each module. This approach enables a standardized evaluation across otherwise diverse module structures and metrics, offering a coherent basis for interpreting relative impact.

This ranking serves as a heuristic tool to compare intervention impact across modules. It is not intended as an absolute measure of effectiveness, but rather as a comparative summary aligned with our research goals.

Findings by Module (Ranked by Effectiveness)

1. **Hot Tag Salience**

   Top-3 Entry: 85.2%
   Score Increase: 96.3%
   A lightweight visual cue ("hot" label) placed on the user's weakest topics resulted

in both widespread score improvement and notable rank shifts.
Conclusion: A highly effective, low-cost intervention that succeeded in changing content priority.

2. **Comment Activation**

   Double Match: 59.3%
   Single Match: 74.1%
   Displaying curated comments at article entry points subtly shaped user behavior, driving meaningful topic reordering.
   Conclusion: Social framing can be a strong motivator of engagement when strategically embedded.

3. **Style Rewriting**

   Top-3 Entry: 47.6%
   Score Increase: 85.7%
   Repurposing low-interest content using a user's preferred writing style helped overcome stylistic barriers and encouraged topic exploration.
   Conclusion: When well-targeted, narrative preference can override topical resistance.

4. **Mixed Content Insertion**

   Top-3 Entry: 40.7%
   Score Increase: 66.7%
   Blending strong and weak topics in a single article offered some improvement, although less dramatic in re-ranking.
   Conclusion: Non-intrusive and effective in moderation, but limited in stronger reorientation.

5. **Explicit Preference Prompt**

   Variance Reduction: 7.4%
   Score Increase in Weakest Share: 66.7%
   Showing users a radar chart of their topic bias raised awareness but had limited impact on behavior.
   Conclusion: Awareness alone is insufficient without compelling motivational elements.

6. **Baseline Preference Identification**

Purpose: Served as the control condition and targeting foundation.
Conclusion: Not designed to induce change, but enabled all other interventions through accurate weak/strong category mapping.

Table 4.12: Result by Module

Module	**Stricter Measurement**	**Looser Measurement**
Hot Tag Salience	85.2%	96.3%
Comment Activation	59.3%	74.1%
Style Rewriting	47.6%	85.7%
Mixed Content Insertion	40.7%	66.7%
Explicit Preference Prompt	7.4%	66.7%
Baseline Identification	—	—

Table 4.13: Module Commentary Summary

Module	**Comments**
Hot Tag Salience	Most effective in changing content priority
Comment Activation	Strong behavioral signal via social context
Style Rewriting	High score lift, moderate rank change
Mixed Content Insertion	Subtle but weaker rank reshuffling
Explicit Preference Prompt	Limited behavior change despite awareness
Baseline Identification	Diagnostic tool only

4.4. Interpretative Analysis: Why Did Users Change (or Not)?

This section delves into the psychological and behavioral mechanisms underlying user responses to each intervention module. By examining the factors that influenced user engagement or resistance, we aim to understand the effectiveness of these interventions in mitigating filter bubble effects.

4.4.1 Hot Tag Intervention: Social Cues and the Fear of Missing Out (FoMO?)

The Hot Tag Intervention demonstrated significant effectiveness in redirecting user attention toward less familiar topics. This outcome may be attributed to the psychological phenomenon known as the Fear of Missing Out (FoMO) [27], where individuals experience anxiety over the possibility of missing rewarding experiences that others might be having.

The "Hot" label likely served as a social cue, signaling the popularity or importance of certain content, thereby triggering users' FoMO and prompting them to engage with topics they might otherwise overlook.

4.4.2 Comment Prompt Activation: The Drive for Expression Over Content Consumption

The Comment Prompt Activation module's success suggests that users are often motivated more by the opportunity to express their opinions than by the content itself. The prompt to comment may have tapped into users' intrinsic desire for social interaction and self-expression, leading them to engage with content outside their usual preferences. This aligns with findings that social media platforms thrive not merely on content consumption but on facilitating user-generated content and interactions.

4.4.3 Style Rewriting: The Influence of Presentation on Engagement

The Style Rewriting intervention highlighted the significant role that presentation style plays in user engagement. By altering the narrative style of content to match users' preferences (e.g., transforming factual reports into opinion pieces), the module increased users' willingness to engage with topics they typically ignored. This underscores the importance of framing and presentation in content delivery, suggesting that stylistic alignment can be as crucial as the content itself in capturing user interest.

4.4.4 Mixed Content Insertion: Enhancing Interest Through Knowledge Integration

The Mixed Content Insertion module's moderate success indicates that integrating familiar elements into unfamiliar content can change user interest. By blending topics of high user interest with less familiar subjects, the intervention leveraged associative learning, making new information more accessible and engaging. This approach reflects educational strategies where new concepts are introduced through their relation to known ideas, facilitating better comprehension and retention.

4.4.5 Explicit Preference Prompt: Awareness Does Not Equate to Behavioral Change

Despite increasing users' awareness of their content consumption patterns, the Explicit Preference Prompt module failed to produce significant behavioral changes. This suggests a disconnect between cognitive awareness and action, where users recognize their filter bubbles but lack the motivation or tools to alter their behavior. Such inertia may be reinforced by the comfort of familiar content and the effort required to seek diverse perspectives. This finding might partly explain the phenomenon of ideas' polarization and self-reinforcement, even when people acknowledge they are enclosed in filter bubbles or echo chambers [28].

In summary, these analyses reveal that subtle cues and alignment with user preferences can effectively nudge users toward more diverse content consumption. However, merely raising awareness of content biases without providing actionable pathways may not suffice to change entrenched behaviors. Future interventions should consider combining awareness with practical strategies to encourage exploration beyond users' established content preferences.

4.5. Why Awareness Alone Doesn't Change Behavior

One of the most unexpected findings of this study emerged from Module 5 (Explicit Preference Prompt). This module aimed to break users' filter bubbles by visualizing their content distribution using a five-dimensional radar chart, updated in real time over 30 testing rounds. The underlying assumption was that once users were made aware of their informational imbalance, they would consciously adjust their reading behavior toward a more balanced knowledge structure.

However, results showed that only 7.4% of participants reduced their score variance, and while 66.7% did increase their weakest-category share, only a small subset achieved both goals. Compared to other modules, this one had the least behavioral impact.

This raised an essential research question:

Why do users, even when explicitly shown their content imbalance, still choose not to diversify their reading habits?

To explore this, we conducted a targeted follow-up interview to probe users' motivations, attitudes, and resistance toward behavioral change in the presence of self-awareness.

Questionnaire Design

The survey consisted of seven multiple-choice questions, designed to explore both cognitive and emotional factors behind users' behavior in response to Module 5.

1. **What was your first impression of the radar chart visualization?**
   - A. It was accurate and matched my habits
   - B. It was confusing and hard to understand
   - C. I didn't really pay attention to it
2. **Did you attempt to adjust your reading choices after seeing the chart?**
   - A. Yes, I tried to balance my content
   - B. No, I mostly followed my interests anyway
   - C. I didn't think it mattered
3. **How important do you think content diversity is in daily news reading?**
   - A. Very important — it helps avoid bias
   - B. Somewhat important, but depth matters more
   - C. Not very important — I prefer to focus on my passions
4. **When shown unfamiliar topics, what do you usually do?**
   - A. I give them a try

B. I usually skip them

C. I only read them if they look exciting

5. **Did the radar chart affect your reading behavior in the long run?**

   A. Yes, it changed how I choose articles

   B. No impact, I read the same as before

   C. I forgot about it after a while

6. **Why do you think you avoid your weakest categories?**

   A. They're boring or not relevant to me

   B. I don't know enough about them

   C. I just don't enjoy them

7. **Do you believe platforms should help users explore diverse content?**

   A. Yes, even if it's outside my comfort zone

   B. Only if it's done gently through what I like

   C. No, I want full control over what I read

Response Summary

Below is a summary of responses from 27 participants (Table 4.14):

Table 4.14: Response Summary for Module 5 Questionnaire

Question	Most Selected Option(s)	No. of Respondents
Q1	A. "It matched my habits"	14
Q2	B. "Followed interests anyway"	16
Q3	B. "Somewhat important"	13
Q4	B. "Usually skip them"	17
Q5	B. "No impact"	15
Q6	A & C tied: "Boring" or "Not enjoyable"	19 each

Question	Most Selected Option(s)	No. of Respondents
Q7	A & B tied: "Yes" or "Only through what I like"	13 each

The survey highlights a gap between awareness and behavior. While users acknowledged the accuracy of the radar chart and the general value of balanced information, most did not adjust their reading habits. The dominant reasons were lack of interest, low perceived relevance, and emotional disengagement from unfamiliar topics.

Users are not necessarily driven by a desire to broaden perspectives. Instead, they are guided by emotional affinity, cognitive fluency, and enjoyment. Even when confronted with clear visual evidence of bias, behavioral inertia prevails.

Chapter 5
Conclusion

5.1. Summary

This study set out to explore three key research questions: (1) how individuals' reading interests can be identified, (2) how those interests might be influenced or altered, and (3) whether individuals can become aware of such biases and take corrective action. Through the design and evaluation of a series of experimental modules—including interest detection mechanisms, targeted interventions, and user-facing feedback systems—this research has demonstrated viable answers to each of these questions.

First, the study successfully derived user reading profiles based on simplified content categorization and behavioral scoring. Second, several modules—particularly those involving stylistic rewriting and social cues—were found to significantly shift user engagement patterns, indicating that content preferences are malleable under certain conditions. Third, while awareness alone proved insufficient to change behavior for most participants, the inclusion of visual feedback and active choice mechanisms showed promise in promoting self-reflection and intentional diversification.

Taken together, these findings offer both empirical insights and practical pathways for addressing filter bubbles in user-centric and context-neutral ways.

5.2. Limitations

This study presents a multi-faceted experimental framework to evaluate filter bubble interventions; however, several limitations constrain the generalizability and interpretability of the findings. These limitations are outlined below:

5.2.1 Limited Sample Size and Lack of Demographic Diversity

The experiments were conducted with a relatively small group of 27 participants. No systematic demographic profiling was performed, meaning variations in age, educational background, political ideology, or digital literacy were not accounted for. This lack of diversity limits the external validity of the results and reduces the study's ability to detect subgroup-specific behavioral patterns.

5.2.2 Artificial Testing Environment

Participants interacted with news articles through a fixed, researcher-designed interface rather than in their natural digital habitats (e.g., social media feeds or news apps). While this design ensured experimental control and data consistency, it may have induced behavior influenced by observation or task expectations. Social desirability bias and reduced ecological validity are possible consequences.

5.2.3 Absence of Complex Multimedia Formats

This study focused exclusively on text-based content, omitting more complex or popular news formats such as images, videos, and interactive media. The exclusion was intentional—aimed at minimizing content generation burden and controlling for variables—but it limits applicability to contemporary media environments where visual stimuli strongly affect engagement. Multimedia content may interact differently with user attention and cognitive framing, making it a valuable target for future investigations.

5.3. Future Directions

Building upon the findings and limitations of this study, several promising directions can be explored to deepen our understanding of filter bubble dynamics and improve intervention strategies.

5.3.1 Expanding Participant Diversity

Future studies should involve larger and more demographically diverse participant pools. Incorporating users from different age groups, political affiliations, educational backgrounds, and media habits would allow for the examination of how individual characteristics interact with intervention effectiveness. This would also support the development of more personalized and targeted anti-bubble strategies.

5.3.2 Testing in Naturalistic Environments

While this study utilized a controlled experimental interface to isolate behavioral signals, future research should aim to embed interventions within real-world news consumption platforms. Deploying experimental modules in actual social media feeds or news aggregators would yield more ecologically valid results and allow for the measurement of sustained behavior over time rather than short-term reactions.

5.3.3 Incorporating Multimedia Elements

Given the dominance of visual content in digital media, future iterations of this research should investigate how images, videos, infographics, and audio influence filter bubble behavior. Exploring whether stylistic rewrites or content mixing retain their effects in multimedia-rich contexts would provide a more complete understanding of attention and persuasion mechanisms.

5.3.4 Longitudinal Behavior Tracking

This study focused on short-term, session-based behavior changes. Future research could employ longitudinal methods to determine whether interventions have durable effects on content preferences and whether users eventually return to their original patterns. Tracking changes across months or even years would better reflect the challenge of breaking long-standing filter bubbles.

5.3.5 Integrating User Agency and Feedback

Beyond passive interventions, future systems might actively engage users in reflecting on their behavior through feedback loops, interactive dashboards, or behavioral nudges. This could help evaluate whether giving users more control and awareness leads to sustained diversification or merely momentary compliance.

References

[1] Eli Pariser. *The filter bubble: What the Internet is hiding from you.* penguin UK, 2011.

[2] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 677–686, New York, NY, USA, 2014. Association for Computing Machinery. URL: `https://doi.org/10.1145/2566486.2568012`, `doi:10.1145/2566486.2568012`.

[3] Jae Kook Lee, Jihyang Choi, Cheonsoo Kim, and Yonghwan Kim. Social Media, Network Heterogeneity, and Opinion Polarization. *Journal of Communication*, 64(4):702–722, 01 2014. URL: `https://doi.org/10.1111/jcom.12077`, `arXiv:https://academic.oup.com/joc/article-pdf/64/4/702/22322442/jjnlcom0702.pdf`, `doi:10.1111/jcom.12077`.

[4] Y. Min, T. Jiang, C. Jin, Q. Li, and X. Jin. Endogenetic structure of filter bubble in social networks. *Royal Society Open Science*, 6(11), 2019. Original work published November 1, 2019. URL: `https://doi.org/10.1098/rsos.190868`, `doi:10.1098/rsos.190868`.

[5] Qazi Mohammad Areeb, Mohammad Nadeem, Shahab Saquib Sohail, Raza Imam, Faiyaz Doctor, Yassine Himeur, Amir Hussain, and Abbes Amira. Filter Bubbles in Recommender Systems: Fact or Fallacy — A Systematic Review, 2023. URL: `https://arxiv.org/abs/2307.01221`, `arXiv:2307.01221`.

[6] Gabriel Machado Lunardi, Guilherme Medeiros Machado, Vinicius Maran, and José Palazzo M. de Oliveira. A metric for Filter Bubble measurement in recommender algorithms considering the news domain. *Applied Soft Computing*, 97:106771, 2020. URL: `https://www.sciencedirect.com/science/article/pii/S1568494620307092`, `doi:https://doi.org/10.1016/j.asoc.2020.106771`.

[7] Reuters Institute for Journalism and Richard Fletcher. *The truth behind filter bubbles: Bursting some myths*. Reuters Institute, 2021.

[8] Gali Einav, Ofir Allen, Tamar Gur, Yossi Maaravi, and Daniel Ravner. Bursting filter bubbles in a digital age: Opening minds and reducing opinion polarization through digital platforms. *Technology in Society*, 71:102136, 2022. URL: `https://www.sciencedirect.com/science/article/pii/S0160791X22002779`, `doi:https://doi.org/10.1016/j.techsoc.2022.102136`.

[9] Tobias D Krafft, Michael Gamer, and Katharina A Zweig. Wer sieht was? Personalisierung, Regionalisierung und die Frage nach der Filterblase in Googles Suchmaschine. *Kaiserslautern: Algorithm Watch. Available at https://bit. ly/3pDcL41 Echo chambers*, 2018.

[10] Denis Parra and Peter Brusilovsky. User-controllable personalization: A case study with SetFusion. *International Journal of Human-Computer Studies*, 78:43–67, 2015.

[11] Thilini Gamage. Realizing the filter bubble: Algorithmic awareness, imagination, and power in information filtering on social media. 01 2025.

[12] Nils Plettenberg, Johannes Nakayama, Poornima Belavadi, Patrick Halbach, Laura Burbach, André Calero Valdez, and Martina Ziefle. User behavior and awareness of filter bubbles in social media. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Communication, Organization and Work: 11th International Conference, DHM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*, pages 81–92. Springer, 2020.

[13] Raymond Nickerson. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 2:175–220, 06 1998. `doi:10.1037/1089-2680.2.2.175`.

[14] Natalie Jomini Stroud. Media Use and Political Predispositions: Revisiting the Concept of Selective Exposure. *Political Behavior*, 30(3):341–366, 2008. URL: `http://www.jstor.org/stable/40213321`.

[15] Drew Westen, Pavel S. Blagov, Keith Harenski, Clint Kilts, and Stephan Hamann. Neural Bases of Motivated Reasoning: An fMRI Study of Emotional Con-

straints on Partisan Political Judgment in the 2004 U.S. Presidential Election. *Journal of Cognitive Neuroscience*, 18(11):1947–1958, 11 2006. URL: https://doi.org/10.1162/jocn.2006.18.11.1947, arXiv:https://direct.mit.edu/jocn/article-pdf/18/11/1947/1935646/jocn.2006.18.11.1947.pdf, doi:10.1162/jocn.2006.18.11.1947.

[16] Charles Dunn, Jacob Hunter, William Steffes, Zackary Whitney, Michael Foss, Jere Mammino, Adam Leavitt, Spencer D Hawkins, Alexander Dane, Martin Yungmann, et al. Artificial intelligence—derived dermatology case reports are indistinguishable from those written by humans: A single-blinded observer study. *Journal of the American Academy of Dermatology*, 89(2):388–390, 2023.

[17] Kruakae Pothong and R Nielsen. *Reuters Institute digital news report 2016: Asia-Pacific supplementary report*. Reuters Institute for the Study of Journalism, 2016.

[18] Ryosuke Homma, Yoshifumi Seki, Mitsuo Yoshida, and Kyoji Umemura. Analysis of short dwell time in relation to user interest in a news application. In *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, pages 219–226, 2020. doi:10.1109/WIIAT50758.2020.00033.

[19] Robert M. Entman. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993.

[20] Amy Mitchell, Jeffrey Gottfried, Michael Barthel, and Nami Sumida. Distinguishing Between Factual and Opinion Statements in the News. https://www.pewresearch.org/journalism/2018/06/18/distinguishing-between-factual-and-opinion-statements-in-the-news/, June 2018. Pew Research Center.

[21] Robert M. Entman. Framing bias: Media in the distribution of power. *Journal of Communication*, 57(1):163–173, 2007. doi:10.1111/j.1460-2466.2006.00336.x.

[22] Graham N. Dixon and Christopher E. Clarke. Heightening uncertainty around certain science: Media coverage, false balance, and the autism-vaccine controversy. *Science Communication*, 35(3):358–382, 2013. URL: https://journals.sagepub.com/doi/10.1177/1075547012458290, doi:10.1177/1075547012458290.

[23] Kimberly Gross. Framing persuasive appeals: Episodic and thematic framing, emotional response, and policy opinion. *Political Psychology*, 29(2):169–192, 2008. URL: https://www.jstor.org/stable/20447111.

[24] Kobie Van Krieken and José Sanders. Framing narrative journalism as a new genre: A case study of the Netherlands. *Journalism*, 18(10):1364–1380, 2017. doi:10.1177/1464884916671156.

[25] Vitalis Torwel. Moral framing and the development of political conflicts: An analysis of new york times' stories in the build up to the 2003 iraqi war. *ATHENS JOURNAL OF MASS MEDIA AND COMMUNICATIONS*, 1:257–274, 09 2015. doi:10.30958/ajmmc.1-4-2.

[26] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1-2):81–93, 1938. URL: https://www.jstor.org/stable/2332226.

[27] M. Akbari, M. Seydavi, S. Palmieri, G. Mansueto, G. Caselli, and M. M. Spada. Fear of missing out (FoMO) and internet use: A comprehensive systematic review and meta-analysis. *Journal of Behavioral Addictions*, 10(4):879–900, 2021. doi:10.1556/2006.2021.00083.

[28] Giacomo Talamanca and Selene Arfini. Through the newsfeed glass: Rethinking filter bubbles and echo chambers. *Philosophy & Technology*, 35, 03 2022. doi:10.1007/s13347-021-00494-z.