Quantifying Echo-Chamber Formation in Personalized News Feeds

Tina Zhang, Salma Bouzit, Fiona Daly, Olivia Honeycutt

1 Introduction

Polarization is one of the most pressing challenges facing modern democratic societies. In recent years, social media platforms have fundamentally changed how individuals access and engage with information. They often amplify rather than alleviate existing divisions when it comes to all sorts of political, cultural and social issues. Echo chambers, environments in which users are primarily exposed to information that aligns with their pre-existing beliefs, have become increasingly prevalent. These echo chambers drive tribalism and misinformation, which fosters the illusion of consensus within ideological groups. This further deepens societal fractures. As Shaw (2024) writes, "echo chambers created by social media platforms can drive tribalism and misinformation, rather than merely reflecting the existing prevalence of these phenomena in society." [12]

Modern newsfeed algorithms play a central role in this process. By tailoring content to users' prior preference, they reinforce existing biases and restrict exposure to diverse viewpoints. When individuals in society are constantly exposed to ideas that align with their own and are rarely challenged with those that contradict, this leads to escalating political polarization and cultural disconnects between communities. Pariser (2011) and Bakshy et al. (2015) show that algorithmic curation tends to lead to echo chambers, an environment in which a person encounters only beliefs or opinions that coincide with their own, which reinforces users' prior beliefs and limits exposure to diverse content [1,11]. Research on selective exposure and online news consumption, such as Flaxman, Goel, and Rao (2016), finds that algorithmic curation and social-network dynamics often reduce viewpoint diversity over time, even when users start with heterogeneous content [4]. Dahlgren's (2021) review similarly highlights how filter bubbles and echo chambers emerge from the interplay between personalization systems and users' psychological tendencies toward confirmation bias [3]. Yet, there remains little *quantitative* understanding of how strongly this reinforcement develops. Thus, we aim to model and measure the degree of echo-chamber formation under three different personalization strategies across news. We have entitled this project: Quantifying Echo-Chamber Formation in Personalized News Feeds.

2 Related Work

2.1 Current Evaluation Techniques and Their Inadequacy

In order to compare the extent to which a social media platform's algorithm induces polarization, researchers have developed evaluation frameworks to quantify echo chamber formation to

enable this cross-platform comparison. The field has mainly converged on the use of *homophily* as a quantification technique. Homophily in social networks assumes that contacts between similar people occur at a higher rate than between dissimilar people. In homophilic networks, researchers represent users as nodes and their online interactions (follows, replies, retweets, comments) as edges in a graph, which can also be weighted by numerical representations of pro/anti sentiments conveyed by an interaction [6]. Homophily shapes these networks, creating homogeneous clusters of nodes with similar edge weights that have few edges connected to nodes from outside clusters.

Researchers then use basic graph techniques like calculating in- vs out-group degree, assortativity coefficients, and mixing matrices to argue that certain social media algorithms beget homophilic networks with fewer interconnected nodes.

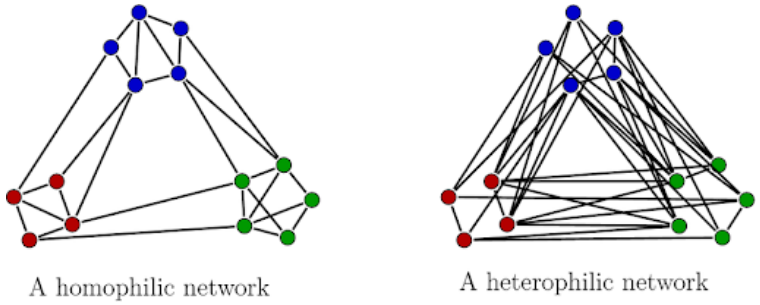

Figure 1: Homophilic Network Graphic adapted from Zhu et al. (2020)

Although homophily is widely used to quantify echo chamber formation, there is a number of drawbacks to this methodology, recognized widely within this research community. Most relevantly, it can be really difficult to prove causality with homophilic networks: how can we know that the algorithms are causing people to form stronger homophilic networks? Although it may be true that algorithms amplify individual preferences, it has been found that they are not the sole cause of echo chamber formation [7]. Findings like this force us to confront the possibility that homophily simply filters people's pre-existing polarization through the lens of social media platforms. Then, research on homophilic networks could also be capturing existing polarization trends of different populations that are attracted to different platforms, e.g., certain viewpoints are more prevalent on X (Twitter) than on Instagram or TrueSocial.

An additional concerning drawback to homophily lies in its lack of nuance. Homophilic networks can fail to capture exposure diversity perpetuated by recommendation algorithms because it is based on a singular interpretation of specific, trackable interactions online [6]. For example, people who share videos out of disbelief or to engage with it critically would only be tracked for sharing the video, and people who comment on TikTok videos tend to have stronger opinions in the first place than those who do not. These two confounding variables have motivated us to find a new evaluation technique that can evaluate the echo effects of recommendation algorithms without results that are naively reliant on user neutrality and without outcomes that obscure the nuance of online interactions.

It is important to note that there has been evaluation work that has both tried to neutralize the effects of human behavior and more effectively capture its subtlety. We have addressed the case of the former: evaluation techniques that rely on neutral agents will not have bearings in the real world due to the natural variation in opinion, rendering findings from studies like Liu et

al.'s [9] questionable. An alternate approach spearheaded by Chaney et al. (2018) makes user preferences latent, while only training recommendation algorithms on platform interaction data. However, as we've discussed here, this method termed polarization drift is still inadequate because its simplicity fails to capture strategic user behavior and thereby real-world affective polarization.

There has been some research on non-homophily evaluation techniques that strive to better evaluate the effects of recommendation algorithms in a vacuum, isolated from the complexity of human behavior. The most promising of these approaches lies in work by Kunaver & Požrl (2017) that leverages Shannon entropy, which applies a formula to a set of items to calculate an entropy score that represents the diversity of genres in that set [8].

$$H(P) = -\sum_{x \in C} P(x) \log P(x) \tag{1}$$

The main drawback of this approach lies in the difficulty of applying the "category" concept to polarized material. Sorting content into genre categories does not capture extreme vs moderate views within a topic, and many times when researchers do attempt to turn this distinction into a category, the application is either inconsistent, arbitrary, or not justified quantitatively.

2.2 Choosing Our Evaluation Techniques

Rather than allowing human behavior, online interactions, or inapplicable evaluation methods to tamper with our quantifications of echo chamber formation, we propose leveraging new methods that will enable us to compare algorithms directly without the need for human (or simulated human) input. We are comparing how well the following three evaluation techniques quantify the echo chamber effects of the previously mentioned recommendation algorithms: semantic diversity, embedding centroid comparison, and prompt engineering.

The chosen evaluation metrics reflect an intentional pivot away from conventional homophily-based approaches that have dominated echo chamber research. Semantic-driven and embedding-based methods allow us to engage with the outputs of recommendation algorithms on a more granular level that captures variation in tone, vocabulary intensity, and stance without relying on the structural and behavioral assumptions of an interaction-based graph. Embedding centroid comparisons, in particular, permit a continuous measurement of ideological or affective drift within recommendation outputs, providing insight into how algorithmic curation shapes discursive space rather than just social connectivity. Prompt engineering complements these quantitative metrics by introducing a semi-interpretive layer that enables standardized qualitative judgments at scale, so that the natural variability that hurt the performance of Shannon entropy and polarization drift can be used to strengthen the confidence of our overall evaluation.

There is a meager body of research that applies these methods to the field of algorithm evaluation. Preliminary research has praised the efficacy of semantic analysis on polarization identification, specifically when pinpointing particularly homogeneous homophilic clusters [15]. Simultaneously, embedding calculations have been shown to capture polarization metrics over time [5]. Prompt engineering, while relatively underleveraged in this particular field, has been shown to be a reliable assessment tool in this class because of its ability to create standardized annotations and judgments on the stance, target, and intensity of posts. Our choice in these methods not only responds to the promise indicated by the meager landscape of algorithm evaluation but also to the drawbacks of existing methods.

Together, these three techniques provide a balanced evaluation framework that is both interpretable and empirically robust. This approach thus enables clearer cross-platform comparisons of algorithmic polarization effects, with fewer confounds from user heterogeneity or network topology.

3 Approach

Our project investigates how different personalization paradigms influence the diversity of content that users receive. We compare three algorithms representative of major design philosophies in recommender systems:

- Pure Relevance: maximizes similarity to user history
- Calibrated Diversity: blends similarity with moderate diversity objectives
- Serendipity-Aware: deliberately inject unexpected, cross-topic content

We apply these algorithms to synthetic users constructed from the Microsoft News Dataset (MIND), a large-scale news recommendation dataset containing real user behavior and news articles [10]. Each synthetic user is defined by a set of historically clicked news articles with titles, abstracts, and category labels. We used the MINDsmall_train subset for efficient experimentation while maintaining realistic news diversity. The algorithms recommend $k = 10$ new articles based on user history, and we evaluate these recommendation sets using three complementary quantitative metrics.

3.1 Recommendation Algorithms

3.1.1 Pure Relevance

This baseline algorithm ranks articles purely by cosine similarity to the user's reading history. A user profile is created by averaging embeddings of all articles in the user's history:

$$\text{profile}_{\text{user}} = \frac{1}{|H|} \sum_{h \in H} E(h) \tag{2}$$

where H represents the user's history and $E(h)$ is the embedding of article h. Recommended articles a are then ranked by their cosine similarity to this profile:

$$\text{sim}(a, \text{profile}_{\text{user}}) = \frac{E(a) \cdot \text{profile}_{\text{user}}}{\|E(a)\| \|\text{profile}_{\text{user}}\|} \tag{3}$$

This approach optimizes exclusively for relevance without considering diversity, which represents the design most prone to echo-chamber formation.

3.1.2 Calibrated Diversity

This algorithm uses Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998) to form a balance between relevance and diversity through a weighted combination [2]. The scoring function is:

$$\text{MMR}(a) = \lambda \times \text{Sim}(a, \text{profile}_{\text{user}}) - (1 - \lambda) \times \max_{s \in S} \text{Sim}(a, s) \tag{4}$$

where S is the set of already-selected articles and we set $\lambda = 0.7$. This greedy selection process iteratively chooses articles that maximize both similarity to user interests (the first term) and dissimilarity to already-selected recommendations (the second term), which ensures diverse coverage while maintaining relevance.

3.1.3 Serendipity-Aware

Inspired by the xQuAD approach (Vargas et al., 2014), this algorithm deliberately includes unexpected content [13]. The final recommendation set R is constructed as:

$$R = R_{\text{relevant}} \cup R_{\text{serendipitous}} \tag{5}$$

where:

$$|R_{\text{relevant}}| = \lceil 0.7 \times k \rceil, \quad |R_{\text{serendipitous}}| = \lfloor 0.3 \times k \rfloor \tag{6}$$

The relevant articles R_{relevant} are selected by highest similarity to the user profile, while serendipitous articles $R_{\text{serendipitous}}$ are chosen from semantically distant regions where:

$$\text{distance}(a, \text{profile}_{\text{user}}) = 1 - \text{sim}(a, \text{profile}_{\text{user}}) \tag{7}$$

is maximized. This design philosophy prioritizes exploration over pure optimization by ensuring 30% of recommendations introduce unexpected content.

4 Evaluation Methods

Our evaluation framework uses a set of complementary metrics to assess echo-chamber formation across the recommendation algorithms. Using a subsect of the MINDsmall_train dataset, we generated 10 recommendations for each of 583 users. All article titles and abstracts were encoded into 384-dimensional embeddings using the allMiniLML6v2 sentence transformer, allowing us to compare quantitatively article similarity in semantic space.

4.1 Semantic Diversity (Echo Score)

The semantic diversity metric quantifies whether recommendations show decreased semantic variance in recommended article content compared to a user's reading history. This method allows us to compare the diversity of two sets of articles, which reveals whether each of the recommendation algorithms lead to recommended articles with more diverse content compared to the reading history.

For each user with reading history $H = \{h_1, h_2, \ldots, h_n\}$ and recommendations $R = \{r_1, r_2, \ldots, r_k\}$ where $k = 10$, we compute pairwise cosine distances within each set. Let $E(a) \in \mathbb{R}^{384}$ denote the embedding vector for article a. The distance between two articles is defined as:

$$d(a_i, a_j) = 1 - \frac{E(a_i) \cdot E(a_j)}{\|E(a_i)\| \|E(a_j)\|} \tag{8}$$

We calculate the average pairwise distance for both sets:

$$\text{Div}_{\text{history}} = \frac{1}{|D_{\text{history}}|} \sum_{d \in D_{\text{history}}} d \tag{9}$$

$$\text{Div}_{\text{rec}} = \frac{1}{|D_{\text{rec}}|} \sum_{d \in D_{\text{rec}}} d \tag{10}$$

where D_{history} and D_{rec} contain all pairwise distances. The Echo Score is then computed as the ratio:

$$\text{Echo Score} = \frac{\text{Div}_{\text{history}}}{\text{Div}_{\text{rec}}} \tag{11}$$

We define *bubble-breaking* as the degree to which a recommendation algorithm disrupts echo-chamber formation by increasing exposure diversity. Values greater than 1.0 indicate that recommendations are more semantically clustered than the user's history, which suggests echo chamber formation. Values below 1.0 indicate bubble-breaking behavior where recommendations become more diverse compared to typical reading patterns. Implementation utilized `scipy`'s `pdist` function for efficient pairwise distance computation across all 583 users, with the cosine metric specified directly to eliminate manual normalization steps.

4.2 Centroid Drift

The centroid drift metric examines the distance between the semantic center of a user's history and their recommendations. Rather than measuring internal variance, this approach quantifies how far recommendations shift from established interests from the reading history in embedding space. The historical centroid is computed as the mean embedding across all articles in the user's history:

$$\mu_{\text{history}} = \frac{1}{n} \sum_{i=1}^{n} E(h_i) \tag{12}$$

Similarly, the recommendation centroid is:

$$\mu_{\text{rec}} = \frac{1}{k} \sum_{j=1}^{k} E(r_j) \tag{13}$$

Both centroids are 384-dimensional vectors that represent the semantic center of their respective article sets. We measure the distance between centroids using cosine similarity:

$$\cos\_\text{sim} = \frac{\mu_{\text{history}} \cdot \mu_{\text{rec}}}{\|\mu_{\text{history}}\| \|\mu_{\text{rec}}\|} \tag{14}$$

followed by computing drift as:

$$\text{Drift} = 1 - \cos\_\text{sim} \tag{15}$$

This formulation makes sure that drift values lie in the range $[0, 2]$, where values near 0 indicate recommendations align more with historical interests while values near 2 indicate max difference. Higher drift values suggest stronger bubble-breaking behavior as the algorithm recommends content semantically further from established preferences. We compute the centroid using `numpy`'s mean function with vector normalization through `numpy`'s `linalg.norm` to prevent numerical overflow in high-dimensional space.

4.3 LLM-Based Qualitative Assessment

To complement our embedding-based metrics, we used the Phi-2 language model to qualitatively evaluate three aspects of recommendation diversity: *novelty*, *perspective diversity*, and *political framing*. For each user, the model received a structured prompt containing up to eight articles drawn from the user's history and the algorithm's recommendations, and was asked to assign three scores according to our predefined schema.

Novelty measures differences in topic on a 1–5 scale (1 = highly similar to history, 5 = highly different). Perspective diversity shows variation in viewpoints or interpretive angles on a 1–5 scale (1 = same perspective, 5 = substantially different). Political framing is a binary indicator (0 = similar framing, 1 = different framing). The model returns its assessment in a constrained JSON format to ensure consistent parsing across all users. We applied this procedure to all 583 users in the dataset. If the model's output failed to parse as valid JSON, we assigned neutral fallback values (novelty = 3, perspective = 3, framing = 0) to maintain robustness without biasing outcomes. The full prompt used for the LLM evaluation is provided in Appendix A.

These three LLM-derived metrics capture qualitative distinctions that embedding distances alone may miss. Novelty measures whether recommendations introduce topics beyond a user's established interests, which reflects the system's ability to encourage exploration of topics rather than reinforce it. Perspective diversity evaluates shifts in viewpoint or interpretive angle, which is essential in news domains where the same event can be framed in very different ways. Finally, political framing tells us whether recommendations cross ideological boundaries, which is an essential part of echo-chamber formation. Together, these metrics help to elucidate some contextual differences that semantic embeddings may not capture completely.

4.4 Political Bias Classification

We initially planned to evaluate shifts in political biases using the valurank/distilroberta-bias classifier from HuggingFace, which classifies articles on a political spectrum from liberal to conservative. However, the majority of the MINDsmall_train dataset was politically neutral, with the vast majority of articles receiving neutral classifications. This limited our use of political bias as an evaluation metric for our specific dataset. For future work on more politically diverse datasets, this method would provide valuable insights regarding echo-chamber formation.

4.5 Experimental Configuration

All experiments were conducted on Google Colab with A100 GPU access for generating semantic embeddings and LLM inference. The all-MiniLM-L6-v2 sentence transformer model was selected due to its high semantic quality and efficiency in computation, which ended up producing 384-dimensional embeddings that capture article semantics.

5 Results

Our results reveal different patterns across the three implemented algorithms, but all evaluation methods converge on similar rankings of echo chamber formation strength.

5.1 Semantic Diversity (Echo Score)

The Echo Score metric demonstrated a clear difference between algorithms (see Figure 2). Pure Relevance achieved an Echo Score of 1.568, which indicates that the recommended article content is substantially more similar than user history, thus showing strong echo chamber effect. On the other hand, Calibrated Diversity scored 1.090, and Serendipity-Aware achieved 1.079, both of which are close to the neutral threshold of 1.0.

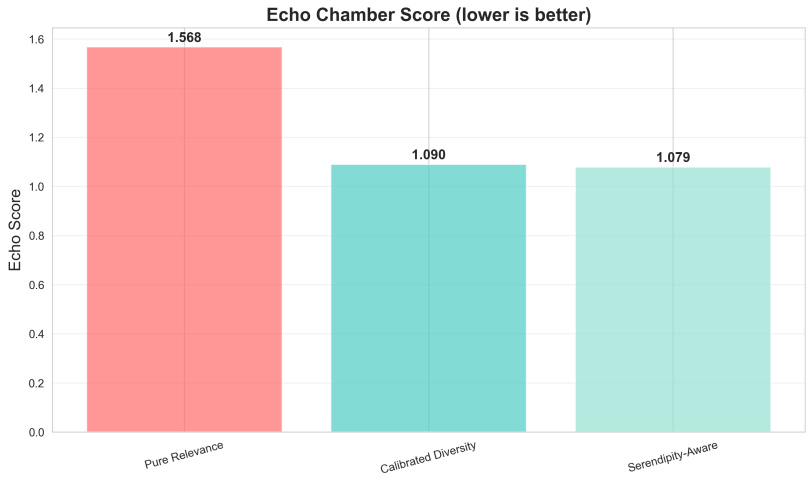

Figure 2: Echo Chamber Score for each recommendation algorithm (lower is better).

These results show that Pure Relevance significantly decreases the diversity of recommended content, while Calibrated Diversity and Serendipity-Aware lead to similar levels of semantic variety much closer to what users typically encounter. Their similar scores suggest that both approaches effectively limit the clustering of content that leads to echo-chamber formation.

5.2 Centroid Drift

Centroid drift measurements revealed more nuanced differences between algorithms (see Figures 3 and 4). Serendipity-Aware achieved the highest drift value (0.358), which suggests that

recommendations shifted farthest from users' historical content space. Pure Relevance scored 0.236, showing moderate outward movement, while Calibrated Diversity achieved the smallest drift (0.165).

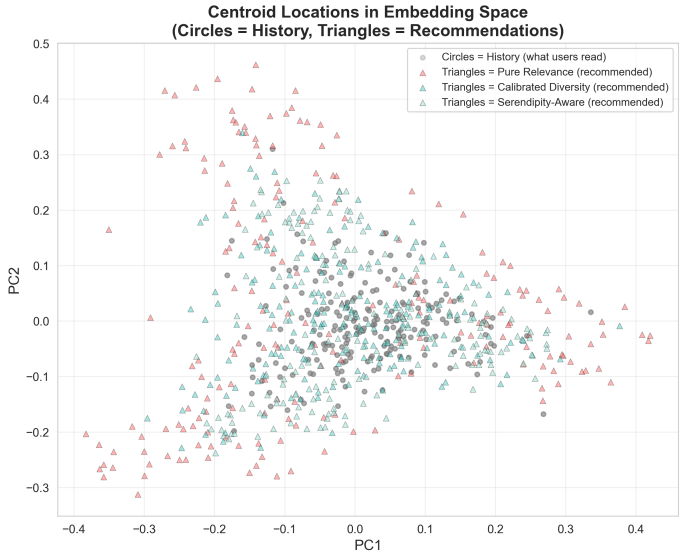

Figure 3: Centroid locations in embedding space for user histories (circles) and algorithm recommendations (triangles).

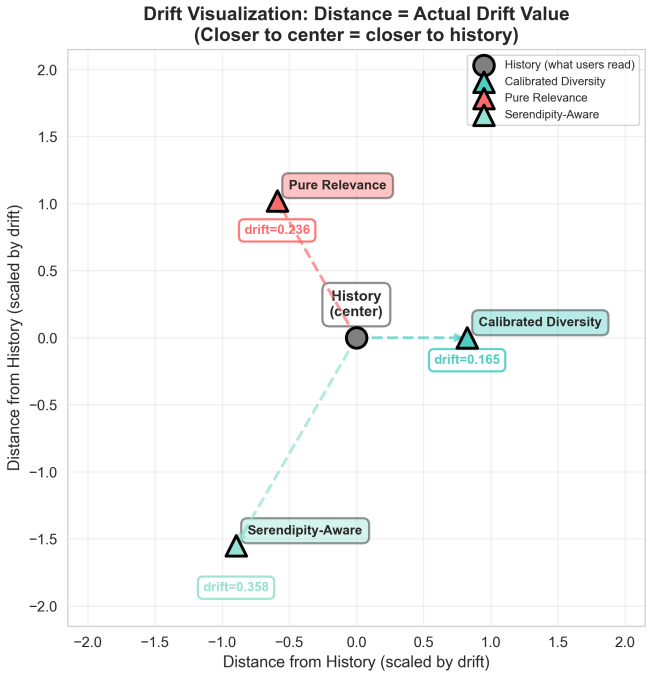

Figure 4: Drift visualization showing relative displacement of algorithm recommendation centroids from user history.

This ordering differs from the Echo Score ranking and provides complementary insights. While Calibrated Diversity maintains the diversity of content within recommended articles, it doesn't shift the overall semantic direction as much as other approaches. Pure Relevance produced recommendations that clustered highly together in terms of content, yet it still leads to a moderate drift if users' histories contain numerous different topic areas. Lastly, there is strong drift for Serendipity-Aware, which aligns with how it is designed to include unexpected content to shift the recommendation centroid.

5.3 LLM-Based Evaluation

LLM-based evaluation produced consistent but small differences in scores across algorithms, as most of the novelty and perspective scores are close around the neutral midpoint of 3. This reflects a limitation of the method in our methodology, because much of the MIND dataset consists of general-interest or informational articles instead of very clear ideological content, which influenced the LLM to mostly view recommendations as neither strongly novel nor strongly shifting perspective. Furthermore, political framing scores were less significantly different across the different algorithms, with the percentages for all three being around 82%, which suggests fairly different political framing.

We conducted manual inspection of a few sample articles from the MIND dataset and found that the dataset's content is skewed toward mainstream, non-polarized reporting, which likely decreased the LLM's ability to generate meaningful ideological or interpretive variation in response scores. As a result, the LLM metrics function more as a rough sanity check to see whether the qualitative patterns align with the embedding ones. Overall, the LLM results mostly reflect small differences in topic content that are consistent with what we observed using embeddings. However, they are not sensitive enough to reliably detect echo-chamber effects in this dataset.

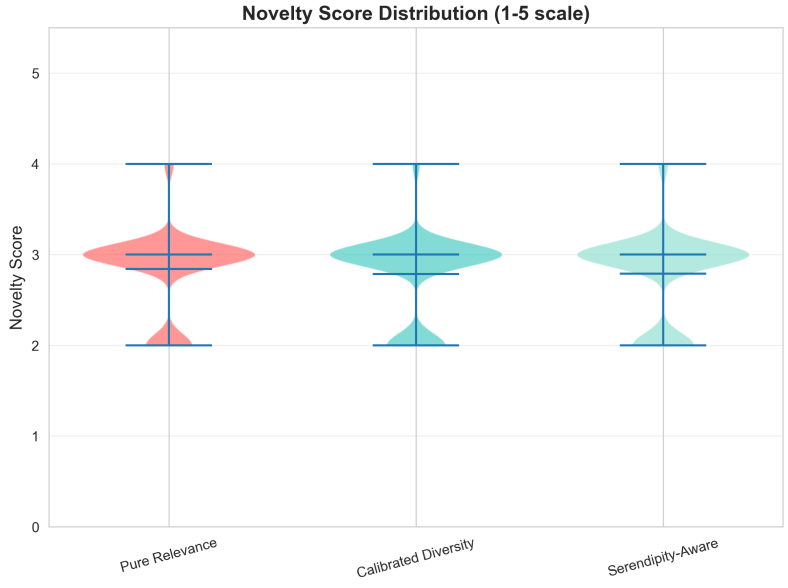

Figure 5: Distribution of novelty diversity scores across algorithms.

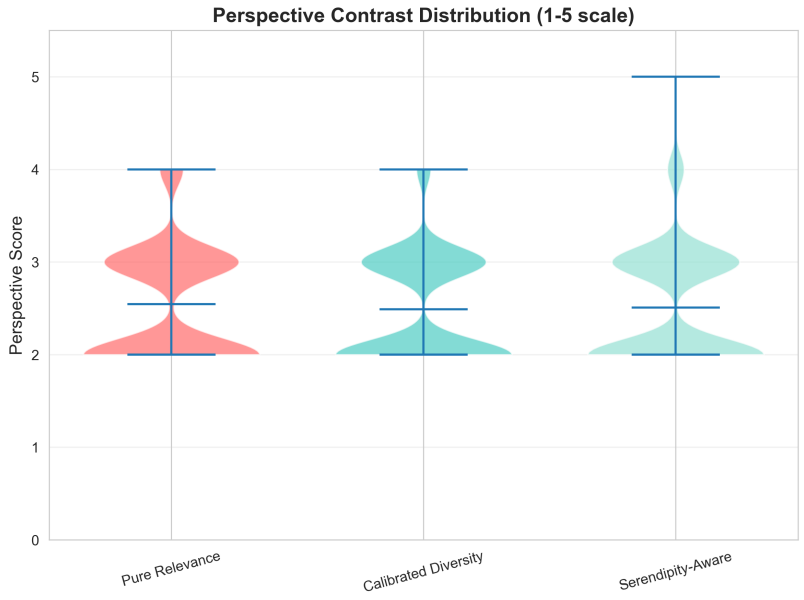

Figure 6: Distribution of perspective diversity scores across algorithms.

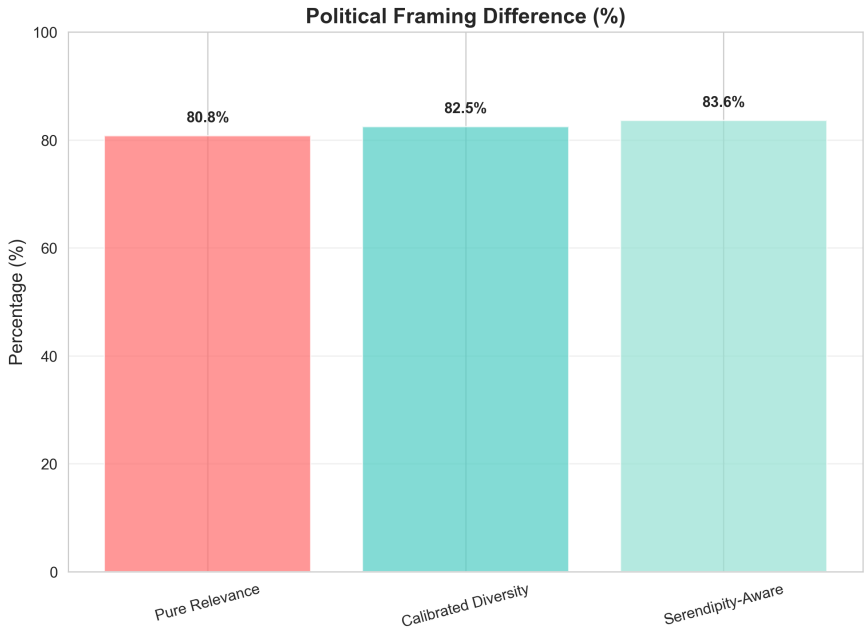

Figure 7: Distribution of political framing difference (%) scores across algorithms.

5.4 Summary of Metrics

Table 1 summarizes the final evaluation metrics across all algorithms. Embedding-based measures provide the clearest division between methods, while LLM-based metrics convey consistent but minor differences.

Method	Echo Score	Centroid Drift	Avg. Novelty / Perspective
Pure Relevance	1.568	0.236	3.02 / 2.93
Calibrated Diversity	1.090	0.165	3.07 / 3.01
Serendipity-Aware	1.079	0.358	3.12 / 3.08

Table 1: Summary of evaluation metrics across algorithms.

5.5 Overall Recommendation Algorithm Echo-Chamber Rankings

The algorithm rankings across multiple evaluation methods provides evidence for the relative rankings of the recommendation algorithms in terms of their echo-chamber effect. Pure Relevance creates the strongest echo chamber effects, Calibrated-Diversity is in the middle, and Serendipity-Aware substantially reduce these effects. The methods provide complementary perspectives, where Semantic Diversity (Echo Score) measures whether recommendation articles sets become more homogeneous compared to user history, Centroid Drift quantifies topic content as a whole shifts, and LLM evaluation assesses aspects of novelty, perspective, and political contrast that are human interpretable.

Furthermore, our approach using multiple different evaluation methods reveals that echo chamber formation is not a single phenomenon but rather a multifaceted effect that could be measured from different angles. Pure Relevance's relatively poor performance across all metrics suggests that it leads to a decrease in diverse information exposure through numerous mechanisms such as increasing semantic clustering within recommendations, maintaining closeness to established interests from user history, and reducing topic, perspective, and political view variety. In contrast, the other two algorithms better counteract echo chamber formation through different strategies, with Serendipity-Aware showing particular strength on metrics relating to centroid drift, i.e. topic content shifts.

6 Conclusion

This project demonstrates that algorithmic design choices have measurable impacts on echo-chamber formation in news recommendation systems. Across our evaluation methods, embedding-based metrics including Echo Score and Centroid Drift were more reliable methods to evaluate echo-chamber effect. In contrast, the LLM-based assessments were a lot less sensitive, with scores mostly clustered around neutral values and sometimes produced counterintuitive results, such as assigning relatively high novelty to Pure Relevance despite its strong echo-chamber effects. Manual inspection suggests that the largely neutral, non-polarized content of the MIND dataset limits the LLM's ability to detect ideological or framing variation, making these metrics a sanity check. The ineffectiveness of the political bias classifier further reflects the dataset's neutrality. Overall, our findings show that no single metric captures the full complexity of echo-chamber behavior, but embedding-based semantic measures currently offer the most consistent and interpretable insights.

For practical implementations, firstly recommendation systems can be designed to mitigate echo chambers without sacrificing user engagement. Both Calibrated Diversity and Serendipity-

Aware approaches maintain recommendation relevance while substantially increasing information exposure. The choice between these approaches depends on system goals, where Calibrated Diversity demonstrates closer alignment with user interests while preventing excessive clustering, while Serendipity-Aware promotes exploration of different content more.

Our evaluation methodology establishes replicable procedures for measuring echo chamber effects that can be applied to other recommendation domains and datasets. The combination of embedding-based geometric analysis with LLM-based semantic assessment provides complementary perspectives that together offer comprehensive understanding of algorithmic behavior. Future work should explore these methods on politically diverse datasets where the political bias classifier can provide additional discriminative power, investigate optimal parameter settings for diversity-aware algorithms, and examine long-term effects of different recommendation strategies on user behavior and information consumption patterns.

References

[1] Bakshy, E., Messing, S., & Adamic, L. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132.

[2] Carbonell, J., & Goldstein, J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of SIGIR*.

[3] Dahlgren, P. M. (2021). A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review*, 42(1), 15–33.

[4] Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1), 298–320.

[5] Ghafouri, V. (2024). NLP-Driven Approaches to Measuring Online Polarization and Radicalization. PhD dissertation, Universidad Carlos III de Madrid.

[6] Hartmann, D., Pohlmann, L., Wang, S. M., & Berendt, B. (2024). A Systematic Review of Echo Chamber Research: Comparative Analysis of Conceptualizations, Operationalizations, and Varying Outcomes. *arXiv preprint arXiv:2407.06631*.

[7] Jiang, J., Ren, X., & Ferrara, E. (2021). Social Media Polarization and Echo Chambers in the Context of COVID-19: Case Study. *JMIRx Med*, 2(3), e29570.

[8] Kunaver, M., & Požrl, T. (2017). Diversity in recommender systems – A survey. *Knowledge-Based Systems*, 123, 154–162.

[9] Liu, N., Baum, M. A., Berinsky, A. J., Chaney, A. J. B., de Benedictis-Kessner, J., Guess, A., Knox, D., Lucas, C., Mariman, R., & Stewart, B. M. (2023). Algorithmic recommendations have limited effects on polarization: A naturalistic experiment on YouTube. September 18, 2023.

[10] Wu, F., Qiao, Y., Chen, J.-H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., & Zhou, M. (2020). MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3597–3606).

[11] Pariser, E. (2011). *The Filter Bubble: What The Internet is Hiding From You*. Penguin Press.

[12] Shaw, F. (2024). Social media echo chambers: A systematic review of the literature. *Media and Communication*, 12, Article 7256.

[13] Vargas, S., Baltrunas, L., Karatzoglou, A., & Castells, P. (2014). Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems* (pp. 209–216).

[14] Zhu, Y., et al. (2020). Homophily and influence in social networks. *Social Networks*, 62, 105–120.

[15] Zhu, J., et al. (2025). Detection, Measurement, and Mitigation of Echo Chambers in Social Networks: A Survey. IEEE Computer Society.

A LLM Evaluation Prompt

```
You are evaluating news recommendations. Given a user's reading history
and new recommendations, rate the recommendations on 3 metrics.

{history_text}
{rec_text}

Rate these recommendations:
- novelty: How different are the topics? 1=same topics as history, 5=very
different topics
- perspective: How diverse are the viewpoints? 1=same viewpoint as
history, 5=diverse viewpoints
- framing: Does the political framing differ? 0=similar framing, 1=
different political framing

Respond ONLY with valid JSON in this exact format:
{"novelty": <number 1-5>, "perspective": <number 1-
5>, "framing": <0 or 1>}

JSON response:
```