AI Review Report

Paper: Attention Inequality on X/Twitter: Evidence from English-Language Posts

Paper ID: 28

Review Service: Reviewer3.com

Generated: May 09, 2026 at 05:26 UTC

Round 1 - Completed

Paper version: v1

Submitted: March 16, 2026 at 14:54 UTC

Completed: March 16, 2026 at 14:58 UTC

reviewer1

The inclusion of the 2026 cross-sectional data point (the red star) in Figure 5 alongside the 2009?2018 EPFL fixed-cohort data visually and logically implies a continuous trend that the underlying data structures cannot support. The 2026 data was collected via a stopword-filtered cross-section of the entire active platform, whereas the EPFL data tracked fixed, legacy-verified cohorts longitudinally. Because the sampling frames, user populations, and platform mechanics (algorithmic vs. chronological feeds) are entirely disjoint between these two datasets, the 2026 retweet Gini cannot logically be used to validate or extend an "upward trend" from the EPFL data. I recommend removing the 2026 data points from the Figure 5 time series and revising the text to treat the historical and contemporary datasets as strictly separate analyses, avoiding causal or continuous trend language between them.

reviewer1

The timeline panel used to decompose variance into between-user and within-user components samples "up to 100 recent original tweets per author." As acknowledged in footnote 6, this means the temporal span covered by the data varies drastically between prolific users (a few days) and infrequent users (months or years). Consequently, the time period of observation is perfectly confounded with user posting frequency. Because engagement dynamics and platform algorithms fluctuate over time, the compressed time window for prolific users likely suppresses their within-user variance artificially, thereby inflating the overall between-user variance estimate (the reported 87% / 81%). To ensure the within-user variance is measured over a valid and comparable baseline, I recommend re-calculating the variance decomposition using a standardized time window for all sampled users (e.g., all tweets posted within a fixed 14- or 30-day period).

reviewer1

The primary bot filter (Criterion i) removes accounts with a mean impressions-to-followers ratio below 0.5%. Because the core regression analysis models the relationship between impressions (the dependent variable) and followers (the independent variable), filtering the dataset based on the direct ratio of these two exact variables introduces a mechanical truncation of the dependent variable's lower tail. While the authors present a sensitivity analysis omitting this filter?which notably drops the quadratic coefficient from 0.053 to 0.028?the primary models and variance decompositions still rely on the truncated data. To avoid hardcoding a performance floor into the dependent variable and artificially inflating the superstar effect, I recommend using the non-performance-based filter (e.g., automation and ghost accounts only) as the primary specification for all main text results, moving the aggressive performance-based filter to the sensitivity analysis.

reviewer1

The study claims to measure attention inequality across English-language posts, but the primary cross-sectional sampling relies on a disjunction of ten specific English stopwords. This method structurally excludes non-text media (images, videos without text) and short-form posts lacking these specific words. Because highly viral content on modern

social media is frequently visual, this sampling strategy is perfectly confounded with content format. The findings logically support conclusions about *text-heavy* English posts, but the strength of the claim exceeds the evidence when generalized to the platform's overall attention economy. I recommend revising the manuscript's claims throughout the abstract and discussion to explicitly scope the conclusions to long-form or text-heavy posts, rather than treating the stopword query as a representative sample of all English-language platform activity.

reviewer2

Residual Overdispersion in the Negative Binomial Model: In Appendix A, the authors use a Negative Binomial GLM to check robustness against the log transformation. The text states that the model accommodates overdispersion and notes a Pearson dispersion statistic of approximately 8. However, a Pearson statistic of 8 for a *fitted* Negative Binomial model indicates severe residual overdispersion that the model's variance function has not fully captured (a well-fitting model should yield a value near 1). The authors should clarify whether the standard errors reported in Table 8 are robust/sandwich standard errors that account for this residual overdispersion. If they are not, the standard errors may be substantially underestimated, and robust standard errors should be applied to ensure the significance of the findings holds.

reviewer2

Post-Treatment Variables in the "Full" Model: In Table 5 and the associated marginal effects analysis (Table 7, Figure 4), the authors include engagement metrics (likes, retweets, quotes) as covariates to model impressions. Because engagement is a direct downstream consequence of impressions (as the authors acknowledge via the "accounting identity"), these variables act as post-treatment covariates. While the authors appropriately disclaim causal inference, interpreting the conditional association of followers *controlling for* engagement (e.g., the interaction terms and marginal effects in Table 7) is statistically problematic, as it conditions on a downstream outcome and risks collider bias. The authors should explicitly discuss the statistical limitations of conditioning on post-treatment variables when interpreting the marginal associations of follower counts in the full model.

reviewer2

Reproducibility of IPW Estimates: The authors report an IPW-corrected between-user variance share of 81%, noting that the weights are based on the ratio of cross-section strata proportions to panel strata proportions across "six log-spaced strata." However, the exact definitions of these strata (e.g., the specific follower count bin edges) and the empirical proportions used to calculate the weights are not reported in the manuscript. To ensure full reproducibility of this key population-corrected estimate, please provide the strata definitions and the corresponding weight calculations, either in the main text or a supplementary table.

reviewer2

Varying Time Windows in Variance Decomposition: The within-user variance decomposition relies on up to 100 recent tweets per user, which spans a median of 14 days but varies widely across users (from a few days to several months, as noted in footnote 6). Because engagement variance may be non-stationary (e.g., subject to temporal autocorrelation or topic shifts over time), calculating within-user variance over vastly different time scales for different users may distort the relative within-user versus between-user variance shares. The authors should consider reporting a robustness check for the variance decomposition that restricts the analysis to a standardized time window (e.g., tweets from the past 14 or 30 days) for all users to ensure the within-user variance is strictly comparable across the sample.

reviewer3

The study applies a rigorous composite filter to remove bot *authors* from the dataset, but it does not address the potential impact of non-human (bot or scraper) *views* on the impression counts themselves. Because impressions (screen renders) are the primary dependent variable, bot-driven inflation of views?which may disproportionately target high-profile accounts or trending topics?could skew the inequality metrics. The authors should discuss this limitation and how unmeasured bot consumption might influence the interpretation of the extreme Gini coefficient.

reviewer3

The temporal analysis (Fact 3) compares a 2009?2018 historical archive to a 2026 cross-sectional snapshot. The 8-year

gap between 2018 and 2026 omits critical periods of platform evolution, making the assertion of a continuous or secular trend highly speculative. The authors should explicitly address the limitations of inferring trends across this missing window, particularly given the major structural and algorithmic changes the platform underwent during that specific timeframe.

reviewer3

The stopword query used to approximate a random sample of the English-language tweet stream ("the OR just OR that OR this OR was OR with OR but OR what OR been OR they") relies heavily on standard English syntax. This approach may inadvertently filter out highly informal speech, short meme-based text, or specific dialects (such as African American Vernacular English) that might omit these specific stopwords in short-form posts. The authors should discuss how this linguistic sampling strategy might bias the representation of certain communities and impact the generalizability of the inequality estimates.

reviewer3

In Section 1, the manuscript compares the 2026 impression Gini (0.965) to YouTube views from 2010 (Cha et al.) and scientific citations from 1976 (de Solla Price). Comparing contemporary, algorithmically-driven feed data to decades-old, structurally different platforms limits the utility of this benchmark. The authors should contextualize this comparison by acknowledging the temporal and structural differences, or ideally, include more contemporary benchmarks for digital attention inequality to better situate the study within the current literature.

reviewer3

The authors rightly note that the observed convex association between followers and impressions could be driven by a mixture of follower-driven ("Following" feed) and algorithm-driven ("For You" feed) regimes, but state that distinguishing these requires unavailable API data. To strengthen the discussion of future directions, the authors could propose specific methodological next steps?such as leveraging user-side data donation, browser extensions, or controlled sock-puppet auditing?that could empirically disentangle these two distribution channels in future research.

Round 2 - Approved

Paper version: v2
Submitted: March 16, 2026 at 15:58 UTC
Completed: March 16, 2026 at 15:58 UTC

Comment 1

Score: 4/4 - Fully addressed

Original comment:
The inclusion of the 2026 cross-sectional data point (the red star) in Figure 5 alongside the 2009?2018 EPFL fixed-cohort data visually and logically implies a continuous trend that the underlying data structures cannot support. The 2026 data was collected via a stopword-filtered cross-section of the entire active platform, whereas the EPFL data tracked fixed, legacy-verified cohorts longitudinally. Because the sampling frames, user populations, and platform mechanics (algorithmic vs. chronological feeds) are entirely disjoint between these two datasets, the 2026 retweet Gini cannot logically be used to validate or extend an "upward trend" from the EPFL data. I recommend removing the 2026 data points from the Figure 5 time series and revising the text to treat the historical and contemporary datasets as strictly separate analyses, avoiding causal or continuous trend language between them.

Author response:
We share the concern about false visual implication of continuity. In v13, we have substantially strengthened the separation. The figure caption now reads: ?The 2026 red star is a separate cross-dataset comparison?not an extension of the EPFL trend: the two datasets differ in sampling frame (fixed cohort vs. cross-section), metric (retweets vs. impressions), and platform era (pre- vs. post-algorithmic feed). The 8-year gap (2018?2026) spans the platform?s most disruptive changes.? The text in Fact 3 now states: ?Our 2026 data is shown as a separate comparison point (red star) but must not be read as extending the EPFL trend: it is a cross-dataset comparison, not a continuous time series.? We

have chosen to retain the 2026 point rather than remove it, because the visual juxtaposition?with appropriate caveats?is more informative to readers than omission. However, the language now makes unambiguously clear that no trend inference is intended.

Reviewer assessment:

The author has fully addressed this concern. While they chose to retain the 2026 data point in Figure 5, they added strong, explicit caveats to both the main text and the figure caption. The revised text and caption now clearly state that the point is a separate cross-dataset comparison, not an extension of the trend, and highlight the differences in sampling, metrics, and the 8-year gap. This effectively prevents the misinterpretation the reviewer was concerned about.

Comment 2

Score: 4/4 - Fully addressed

Original comment:

The timeline panel used to decompose variance into between-user and within-user components samples "up to 100 recent original tweets per author." As acknowledged in footnote 6, this means the temporal span covered by the data varies drastically between prolific users (a few days) and infrequent users (months or years). Consequently, the time period of observation is perfectly confounded with user posting frequency. Because engagement dynamics and platform algorithms fluctuate over time, the compressed time window for prolific users likely suppresses their within-user variance artificially, thereby inflating the overall between-user variance estimate (the reported 87% / 81%). To ensure the within-user variance is measured over a valid and comparable baseline, I recommend re-calculating the variance decomposition using a standardized time window for all sampled users (e.g., all tweets posted within a fixed 14- or 30-day period).

Author response:

This is an important concern that we address directly with new computations. In v13, we restrict each user to tweets within a standardized 14-day (and 30-day) window from their most recent tweet and recompute the variance decomposition: ... The between-user share is virtually identical?in fact, slightly higher under the 14-day window (88.5% vs. 87.8%)?confirming that the compressed time window for prolific users does not drive the result. This is reported in the Fact 1 paragraph with full methodology.

Reviewer assessment:

This comment is fully addressed. The author performed the requested robustness check by re-calculating the variance decomposition using standardized 14-day and 30-day windows. They added text to the manuscript (Section 4.1) reporting that the results are virtually identical to the original analysis, thereby demonstrating that the variable time window did not artificially inflate the between-user variance share. This new analysis completely resolves the concern.

Comment 3

Score: 3/4 - Well acknowledged

Original comment:

The primary bot filter (Criterion i) removes accounts with a mean impressions-to-followers ratio below 0.5%. Because the core regression analysis models the relationship between impressions (the dependent variable) and followers (the independent variable), filtering the dataset based on the direct ratio of these two exact variables introduces a mechanical truncation of the dependent variable's lower tail. While the authors present a sensitivity analysis omitting this filter?which notably drops the quadratic coefficient from 0.053 to 0.028?the primary models and variance decompositions still rely on the truncated data. To avoid hardcoding a performance floor into the dependent variable and artificially inflating the superstar effect, I recommend using the non-performance-based filter (e.g., automation and ghost accounts only) as the primary specification for all main text results, moving the aggressive performance-based filter to the sensitivity analysis.

Author response:

We take this concern seriously and have made the automation-only filter substantially more prominent in v13. Table 3 (sensitivity analysis) now includes a dedicated row for the automation-only filter (criterion iii only; 111 accounts, n = 10,881; Gini = 0.967, Top 1% = 76.4%). The Fact 2 robustness paragraph explicitly discusses the automation-only result (??2 = 0.030, $p < 0.001$) as ?a filter that does not depend on impression metrics at all.? We respectfully retain the full

composite filter as the primary specification, for two reasons. First, the study population is defined as ?accounts that reach a human audience?... Second, both filter specifications confirm the core finding...

Reviewer assessment:

The author has well acknowledged this comment. While they did not adopt the reviewer's recommendation to use the non-performance-based filter as the primary specification, they provided a clear scientific rationale for their decision. More importantly, they made the results from the non-performance-based filter much more prominent by adding a dedicated row for it in the sensitivity table (Table 5) and explicitly discussing its results in the main text, directly addressing the reviewer's underlying concern about circularity.

Comment 4

Score: 4/4 - Fully addressed

Original comment:

The study claims to measure attention inequality across English-language posts, but the primary cross-sectional sampling relies on a disjunction of ten specific English stopwords. This method structurally excludes non-text media (images, videos without text) and short-form posts lacking these specific words. Because highly viral content on modern social media is frequently visual, this sampling strategy is perfectly confounded with content format. The findings logically support conclusions about *text-heavy* English posts, but the strength of the claim exceeds the evidence when generalized to the platform's overall attention economy. I recommend revising the manuscript's claims throughout the abstract and discussion to explicitly scope the conclusions to long-form or text-heavy posts, rather than treating the stopword query as a representative sample of all English-language platform activity.

Author response:

We agree and have strengthened the scope language throughout v13. The abstract now uses ?text-containing posts.? The data description specifies ?text-containing tweets.? The Conclusion now reads ?English-language, text-containing posts on X/Twitter.? The Limitations paragraph explicitly notes that the stopword query ?may underrepresent highly informal speech, short meme-based text, or specific dialects. . . that may omit these particular function words in short-form posts.? We also note that because visual content is often higher-engagement, our inequality estimates are likely conservative.

Reviewer assessment:

This comment is fully addressed. The author has systematically revised the manuscript to explicitly scope their findings to "text-containing posts," as requested. This change is reflected in the abstract, data description, and conclusion, and the limitations of the sampling method are now discussed in more detail. This accurately aligns the paper's claims with its methodology.

Comment 5

Score: 4/4 - Fully addressed

Original comment:

Residual Overdispersion in the Negative Binomial Model: In Appendix A, the authors use a Negative Binomial GLM to check robustness against the log transformation. The text states that the model accommodates overdispersion and notes a Pearson dispersion statistic of approximately 8. However, a Pearson statistic of 8 for a *fitted* Negative Binomial model indicates severe residual overdispersion that the model's variance function has not fully captured (a well-fitting model should yield a value near 1). The authors should clarify whether the standard errors reported in Table 8 are robust/sandwich standard errors that account for this residual overdispersion. If they are not, the standard errors may be substantially underestimated, and robust standard errors should be applied to ensure the significance of the findings holds.

Author response:

An excellent catch. In the original table, we reported model-based SEs, which were indeed underestimated given the residual overdispersion. In v13, Table 8 (Appendix A) now reports HC1 sandwich SEs throughout. The robust SEs are 2?3× larger than the model-based SEs (e.g., SE for (log F)2: 0.009 robust vs. 0.003 default). Crucially, all key coefficients remain highly significant... The Appendix A text now explains: ?The Pearson dispersion statistic is approximately 8, indicating substantial residual overdispersion. . . We therefore report HC1 sandwich standard errors

throughout.?

Reviewer assessment:

The author has fully addressed this critical statistical point. They acknowledged the error, re-estimated the model using robust (HC1 sandwich) standard errors to account for the residual overdispersion, and updated the appendix table (now Table 9) and accompanying text accordingly. By confirming that the key coefficients remain significant even with the corrected, larger standard errors, they have validated the robustness of their finding.

Comment 6

Score: 4/4 - Fully addressed

Original comment:

Post-Treatment Variables in the "Full" Model: In Table 5 and the associated marginal effects analysis (Table 7, Figure 4), the authors include engagement metrics (likes, retweets, quotes) as covariates to model impressions. Because engagement is a direct downstream consequence of impressions (as the authors acknowledge via the "accounting identity"), these variables act as post-treatment covariates. While the authors appropriately disclaim causal inference, interpreting the conditional association of followers *controlling for* engagement (e.g., the interaction terms and marginal effects in Table 7) is statistically problematic, as it conditions on a downstream outcome and risks collider bias. The authors should explicitly discuss the statistical limitations of conditioning on post-treatment variables when interpreting the marginal associations of follower counts in the full model.

Author response:

We agree that this deserves explicit treatment. In v13, we have expanded the endogeneity caveat paragraph (Section 3.3) into an ?Endogeneity and post-treatment bias caveat? that now reads: ?Because engagement is a downstream consequence of impressions, including it as a covariate creates a post-treatment conditioning problem... This risks collider bias... For this reason, we emphasize that the followers-only model (columns 1?2)... provides the cleanest estimate of the follower?impression association... The full model is presented to characterize the equilibrium structure... not to estimate causal effects...?

Reviewer assessment:

This comment is fully addressed. The author added a new, detailed paragraph in Section 3.3 titled "Endogeneity and post-treatment bias caveat." This new text explicitly discusses the problem of conditioning on post-treatment variables, mentions the risk of collider bias, and correctly frames the followers-only model as the cleanest estimate for that specific association, fully clarifying the statistical limitations as requested.

Comment 7

Score: 4/4 - Fully addressed

Original comment:

Reproducibility of IPW Estimates: The authors report an IPW-corrected between-user variance share of 81%, noting that the weights are based on the ratio of cross-section strata proportions to panel strata proportions across "six log-spaced strata." However, the exact definitions of these strata (e.g., the specific follower count bin edges) and the empirical proportions used to calculate the weights are not reported in the manuscript. To ensure full reproducibility of this key population-corrected estimate, please provide the strata definitions and the corresponding weight calculations, either in the main text or a supplementary table.

Author response:

We have added Table 6 to the Data section, reporting the full strata definitions (log-spaced by follower count with boundaries at 102, 103, 104, 105, 106), cross-section counts and proportions, panel counts and proportions, and the resulting IPW weights. This enables full reproducibility of the 81% IPW-corrected estimate.

Reviewer assessment:

The author has fully addressed this comment by adding a new table (Table 2 in the revised manuscript) that provides all the necessary information to reproduce the IPW calculation. The table clearly defines the follower strata and lists the counts, proportions, and final weights. This addition makes a key analysis step completely transparent and reproducible.

Comment 8

Score: 4/4 - Fully addressed

Original comment:

Varying Time Windows in Variance Decomposition: The within-user variance decomposition relies on up to 100 recent tweets per user, which spans a median of 14 days but varies widely across users (from a few days to several months, as noted in footnote 6). Because engagement variance may be non-stationary (e.g., subject to temporal autocorrelation or topic shifts over time), calculating within-user variance over vastly different time scales for different users may distort the relative within-user versus between-user variance shares. The authors should consider reporting a robustness check for the variance decomposition that restricts the analysis to a standardized time window (e.g., tweets from the past 14 or 30 days) for all users to ensure the within-user variance is strictly comparable across the sample.

Author response:

Addressed under R1-2 above. The 14-day and 30-day standardized-window results are now reported in the Fact 1 paragraph and confirm the original estimates.

Reviewer assessment:

This comment is fully addressed. As with the identical comment from Reviewer 1, the author performed the requested robustness check using standardized 14- and 30-day windows. They added text to Section 4.1 showing that the results are robust, which completely resolves the concern about the potential confounding effect of variable time windows.

Comment 9

Score: 4/4 - Fully addressed

Original comment:

The study applies a rigorous composite filter to remove bot *authors* from the dataset, but it does not address the potential impact of non-human (bot or scraper) *views* on the impression counts themselves. Because impressions (screen renders) are the primary dependent variable, bot-driven inflation of views?which may disproportionately target high-profile accounts or trending topics?could skew the inequality metrics. The authors should discuss this limitation and how unmeasured bot consumption might influence the interpretation of the extreme Gini coefficient.

Author response:

This is a valid limitation we had not explicitly discussed. In v13, the Limitations paragraph now includes: ?Bot views: Our filter targets bot authors, not bot viewers. Because impressions count screen renders, bot-driven scraping or automated viewing could inflate impression counts, potentially disproportionately for high-profile accounts or trending topics. If bot views are correlated with account size, the measured inequality could be partly an artifact of non-human consumption...?

Reviewer assessment:

The author has fully addressed this comment. They added a new paragraph to the Limitations section specifically discussing the distinction between bot authors and bot viewers. This new text acknowledges that unmeasured bot consumption could inflate impression counts and potentially skew the inequality metrics, which was the exact concern raised by the reviewer.

Comment 10

Score: 4/4 - Fully addressed

Original comment:

The temporal analysis (Fact 3) compares a 2009?2018 historical archive to a 2026 cross-sectional snapshot. The 8-year gap between 2018 and 2026 omits critical periods of platform evolution, making the assertion of a continuous or secular trend highly speculative. The authors should explicitly address the limitations of inferring trends across this missing window, particularly given the major structural and algorithmic changes the platform underwent during that specific timeframe.

Author response:

We agree. In v13, the Fact 3 text now explicitly states: ?Moreover, the 8-year gap between the EPFL endpoint (2018) and our 2026 snapshot spans the platform?s most consequential structural changes?the 2022 acquisition, the algorithmic open-sourcing, and the transition to paid verification?making any interpolation between the two periods highly speculative. We include the 2026 point only for visual context, not to validate or extend the EPFL trend.? The

figure caption similarly flags the gap. See also the response to R1-1.

Reviewer assessment:

This comment is fully addressed. The author added explicit text to the manuscript (Section 4.3) acknowledging that the 8-year gap between datasets was a period of major platform change and that interpolation is highly speculative. The revised text and figure caption now clearly frame the 2026 data point as a contextual comparison rather than an extension of a trend, resolving the reviewer's concern.

Comment 11

Score: 4/4 - Fully addressed

Original comment:

The stopword query used to approximate a random sample of the English-language tweet stream ("the OR just OR that OR this OR was OR with OR but OR what OR been OR they") relies heavily on standard English syntax. This approach may inadvertently filter out highly informal speech, short meme-based text, or specific dialects (such as African American Vernacular English) that might omit these specific stopwords in short-form posts. The authors should discuss how this linguistic sampling strategy might bias the representation of certain communities and impact the generalizability of the inequality estimates.

Author response:

In v13, the Limitations paragraph now reads: ?The stopword list relies on standard English syntax and may underrepresent highly informal speech, short meme-based text, or specific dialects (e.g., African American Vernacular English) that may omit these particular function words in short-form posts.? This is an inherent limitation of our sampling strategy that we cannot resolve without direct firehose access. We note that the bias is toward underrepresenting communities whose attention dynamics may differ from the text-heavy mainstream sample, and we scope our claims accordingly.

Reviewer assessment:

The author has fully addressed this comment. They added a sentence to the Limitations section that explicitly acknowledges that the stopword sampling strategy may underrepresent informal speech and specific dialects, using the reviewer's example of African American Vernacular English. This addition directly incorporates the reviewer's concern into the discussion of the study's limitations.

Comment 12

Score: 4/4 - Fully addressed

Original comment:

In Section 1, the manuscript compares the 2026 impression Gini (0.965) to YouTube views from 2010 (Cha et al.) and scientific citations from 1976 (de Solla Price). Comparing contemporary, algorithmically-driven feed data to decades-old, structurally different platforms limits the utility of this benchmark. The authors should contextualize this comparison by acknowledging the temporal and structural differences, or ideally, include more contemporary benchmarks for digital attention inequality to better situate the study within the current literature.

Author response:

We have added a Wikipedia page views benchmark (Gini ?0.85; ?) and explicitly acknowledged the temporal limitation: ?We note that these benchmarks span different eras and platform architectures?the YouTube estimate is from 2010 data and the citations benchmark from the 1970s; contemporary algorithmic platforms may have substantially different concentration levels.? Finding strictly contemporary, comparable Gini estimates for other major platforms remains difficult because most platforms do not expose impression data.

Reviewer assessment:

This comment is fully addressed. The author both added a more recent benchmark (Wikipedia page views) and inserted a sentence in the Introduction explicitly acknowledging the temporal and architectural differences between their data and the benchmarks. This provides the necessary context that the reviewer requested, strengthening the comparison.

Comment 13

Score: 4/4 - Fully addressed

Original comment:

The authors rightly note that the observed convex association between followers and impressions could be driven by a mixture of follower-driven ("Following" feed) and algorithm-driven ("For You" feed) regimes, but state that distinguishing these requires unavailable API data. To strengthen the discussion of future directions, the authors could propose specific methodological next steps?such as leveraging user-side data donation, browser extensions, or controlled sock-puppet auditing?that could empirically disentangle these two distribution channels in future research.

Author response:

In v13, the Fact 2 discussion of feed-source composition now concludes: ?Promising methodological avenues include user-side data donation studies (where participants export their feed composition), browser extension-based auditing of feed content, and controlled sock-puppet accounts that vary follower counts while holding content constant. Each approach has limitations (self-selection, Terms of Service compliance, ecological validity), but together they could empirically disentangle the feed-source mixture.?

Reviewer assessment:

The author has fully addressed this suggestion. They expanded the discussion at the end of Section 4.2 to include several specific, well-considered methodological proposals for future research aimed at disentangling feed sources. This addition strengthens the paper by turning a stated limitation into a concrete agenda for future work, as the reviewer recommended.