Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Coverage, Not Volume: Measurement and Reporting Artifacts in Multi-Platform Social-Media Interrupted Time-Series Analyses of COVID-19 Infodemiology

Sui Qin
Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology
Yu Chen
Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology
La Xie
xiela0504@163.com

Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology

Research Article

Keywords: infoveillance, interrupted time series, social media, COVID-19, coverage bias, reporting standards, sentiment analysis, China

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11021884/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Social-media infoveillance advertises “multi-platform” and “longitudinal” coverage, yet timestamp provenance, monthly coverage completeness, and multiple-comparison handling are rarely described—choices that decide whether a policy effect is real or artifactual.

Methods

We retrieved 58,750 verified Chinese COVID-19 posts across five platforms (Weibo, Douyin, Bilibili, Zhihu, Xiaohongshu) via retrospective archival queries (2019-07 to 2026-07). A naive pooled interrupted time-series (ITS) was fit first, then a disclosed post hoc eligibility rule (≥ 80% non-zero monthly coverage; ≥12 months each side of the Nov-2022 interruption). Qualifying platforms entered López-Bernal segmented regression with Newey–West HAC standard errors, Benjamini–Hochberg FDR (m = 2 primary; m = 8 sensitivity), and robustness checks (per-period composition, keyword-yield auditing on terms with ≥ 8 posts per period, leave-one-month-out, placebo breakpoints).

Results

Naive pooled ITS gave a spurious + 8,225.7 posts (P=.03), driven by incomparable pooling, a single Weibo scrape month (44,228 posts, 75.3%), and a composition flip at the interruption (pre 96.1% Bilibili+Zhihu; post 79.8% Weibo—Simpson’s paradox). After eligibility, only Zhihu qualified (Douyin’s breakpoint-density selection defeated identification); no Zhihu coefficient was policy-attributable (smallest q=.064, m = 2; .157, m = 8). The largest Zhihu level (+ 65.5) reflected period-biased vocabulary: 7 post-pandemic terms yielded 25 vs 1,089 posts versus 1.43× for 10 stable terms; it decomposes as + 65.54 = + 20.35 + 45.20, and the excluded term ranks 1/57 of candidate breaks yet is observationally equivalent to topic emergence. Douyin’s slope showed the same artifact (+ 4.78 = + 0.34 + 4.44; 92.9% from terms barely retrieved pre-breakpoint). Sentiment agreement was unusable (κ = 0.12).

Conclusions

Multi-platform ITS validity rests on temporal completeness, timestamp provenance, and stable retrieval yield—not corpus size. We propose six reporting standards and two prospective requirements (freeze and version-control the keyword list before crawling; store crawl_time and publish_time separately, ≥ 12 months comparable coverage per side). Without a pre-registered frozen vocabulary, ITS on platform search output cannot separate instrument change from topic emergence; single-platform restriction is insufficient.

1. Introduction

1.1 The rise of infodemic and social-media infoveillance

The COVID-19 pandemic intensified scholarly interest in the "infodemic"[1] — the overabundance of information, both accurate and misleading, that accompanies health emergencies [2–4]. Social-media infoveillance — harvesting user-generated content to track public attention, sentiment, and thematic concern — has been promoted as a low-cost, high-timeliness complement to traditional surveillance [2, 3, 4], yet systematic reviews of social-media disease surveillance show that core analytical procedures are frequently under-reported [5].

1.2 The unexamined assumption of multi-platform designs

A common design selling point is "more platforms = more robust." This assumption is rarely interrogated on its own terms. Platforms differ radically in (a) when and how often content is scraped, (b) whether the recorded timestamp is a true publication date or a scrape-batch timestamp, and (c) the completeness of monthly coverage. When these dimensions differ by orders of magnitude, pooling platforms into a single longitudinal series is not a strength — it is a confound [6–7].

1.3 Positioning of this study

This manuscript is **not** a policy-effect discovery paper. It is a **methodological cautionary study and a reporting-standards proposal**, building on the social-data bias taxonomy of Olteanu et al. [8] and aligning with the RECORD-PE reporting extension [9]. We use a real, independently verified corpus to demonstrate, with concrete numbers, how coverage screening and multiple-comparison correction reshape — and in places overturn — the conclusions one would draw from a naive analysis. All effect sizes reported here are illustrative of method, not of any causal claim about policy.

1.4 A taxonomy of the artifacts: two analysis layers

The four artifacts are best read not as four independent faults but as stages of one data-generation chain, each acting on a different construct. They fall under two analysis layers (not four independent causal stages): **between-platform** — layer 1, which also produces the platform-composition flip / Simpson's paradox — and **within-platform** — layers 2–4. Layers 1 and 3 share a common Weibo 2022-11 root cause: they are distinct failure modes (between- vs within-platform) of one design limitation (retrospective, unfrozen retrieval). Full details are provided in Table 1.

Table 1

Taxonomy of the four artifacts under two analysis layers. Note. Each artifact acts on a distinct construct. The diagnostic and remedy columns indicate whether the layer alone is sufficient to overturn a naive pooled conclusion. Layer 1 (between-platform) also produces the platform-composition flip / Simpson's paradox; layers 2–4 are within-platform.

Stage	Analysis layer	Construct	Artifact	Diagnostic	Remedy	Alone sufficient?
1	Between-platform	Sampling frame	Coverage pooling (→ composition flip / Simpson's paradox)	Monthly coverage % per platform	Eligibility rule	No — defeats naive pooled inference, reveals further layers
2	Within-platform	Measurement instrument	Sentiment lexicon override	% overwritten; κ vs proxy	Replace instrument + human labels	No — generates false negatives (measurement failure)
3	Within-platform	Retrieval instrument	Period-biased vocabulary	Dispersion of per-term yield ratios r_k (descriptive; the external benchmark ρ is not estimable from a topic-restricted corpus)	Bound by exact decomposition (not point-identified)	No — but principal for the within-platform unidentifiable remainder (our identification-limit claim)
4	Within-platform	Estimand / selection	Breakpoint-density selection	Pre/post median likes; near-zero sample	None in present corpus	**Yes — decisive for rejecting Douyin's coefficient as a policy effect**

Two consequences follow. First, layers 1 and 3 are not independent: both are rooted in Weibo's single 2022-11 scrape, so the "pooled" series fails twice for one underlying cause. Second, only the breakpoint-density selection effect (layer 4) is sufficient on its own to nullify a naive conclusion; layers 1–3 are necessary but not sufficient, and layer 2 is a false-negative generator rather than a false positive. The reporting standards in Section 4.1 therefore target the whole chain, not any single layer.

2. Methods

This observational infoveillance study follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) reporting guidelines; a completed STROBE 2009 checklist is provided as Supplementary Material S10.

2.1 Data provenance and timestamp semantics

The corpus comprises 58,750 publicly available Chinese-language posts on COVID-19 and respiratory disease. They were gathered in 2026 through retrospective archival API queries covering 2019-07 to 2026-07. Each raw crawl product (data/raw/*.jsonl) records two separate time fields for every post: pub_time, the publication date, and crawl_time, the date it was retrieved. We checked pub_time against the raw records by post_id. Across all five platforms the analysis-ready field matches the raw pub_time in 100% of cases (and matches crawl_time in 0%), with no nulls, no placeholder defaults such as 2020-01-01, no future dates, and a span from 2019 to 2026.

Topic-restricted by construction. The corpus is not a sample of platform content; it is the output of a fixed COVID-related query list. Every record carries the query term that retrieved it, and the union across platforms is 17 terms. Query-list breadth is highly unequal across platforms — Zhihu 17 terms, Douyin 9, Weibo 3, Xiaohongshu 2, and Bilibili a single term ("COVID") accounting for all 5,202 of its posts. Two consequences run through the rest of this paper. First, no query-free platform volume exists anywhere in these data, so retrieval-instrument stability cannot be normalised against a platform-level benchmark (Section 2.4). Second, "platform volume" in every series below means *the retrieval yield of this query list on that platform*, and we use the longer phrase wherever the distinction matters.

Critical disclosure. The analysis-ready pub_time *is* a genuine publication date, present and clean for every platform. Full details are provided in Table 2. The genuine provenance limitation in this corpus is twofold: (i) **cross-platform coverage heterogeneity** — publication-date coverage differs by orders of magnitude across platforms (Weibo 99.1% of posts fall in 2022; Xiaohongshu 90% in 2026; only Zhihu spans 2019–2026), so comparable longitudinal coverage exists for only two platforms; and (ii) the crawl_time (scrape-batch) field, although present in the raw JSONL, was **not joined into the analysis-ready CSV** — a provenance gap we now flag, rather than a missing publication date. Both limitations motivate reporting standard (i) and prospective recommendation (b).

Table 2

Platform temporal coverage and timestamp provenance. Note. Posts (n) is the number of retrieved posts; distinct pub dates is the count of distinct publication dates. The crawl-batch timestamp (crawl_time) was present in the raw JSONL but not joined into the analysis-ready CSV (see Section 2.1).

Platform	Analysis-CSV temporal field	Semantics	Non-null %	Posts (n)	Distinct pub dates	Year spread	Coverage note
Zhihu	pub_time	Publication date	100%	2,827	1,209	2019–2026	spread across window
Douyin	pub_time	Publication date	100%	5,946	908	2020–2026	sparse pre-2022
Weibo	pub_time	Publication date	100%	44,626	33	2020, 2022, 2026	99.1% (44,228) in 2022-11; 87.0% (38,843) in 2022-11-01–11 − 10 — a pre-relaxation ~ 10-day cross-section, not a longitudinal series
Bilibili	pub_time	Publication date	100%	5,202	281	2019–2020, 2023, 2026	two clusters
Xiaohongshu	pub_time	Publication date	100%	149	64	2022–2026	90% in 2026
(all)	$crawl_time$ *[raw only]*	Scrape-batch timestamp	—	—	—	uniform ~ 2026-08-08	**not joined to analysis CSV**

2.2 Eligibility rule for interrupted time-series analysis

The eligibility rule is stated openly and is post hoc: a platform qualifies for ITS inference only when it has at least 80% non-zero monthly coverage across the common window 2019-07 to 2026-07, at least 50% non-zero months before the interruption, and at least 12 months on each side of the November-2022 policy break. It was not pre-registered. Threshold sensitivity (50/60/70/80/90%) appears in the Supplementary Material.

Applying the rule to the five platforms:

Table 3

Eligibility assessment of the five platforms against the post-hoc rule. Note. The rule requires ≥ 80% non-zero monthly coverage and ≥ 12 months before and after the November-2022 interruption, evaluated on the common 85-month window (2019-07 to 2026-07). Bilibili, Xiaohongshu and Weibo fall far below the 80% threshold even before the per-segment test.

Platform	Posts, n	Real months with ≥ 1 post (of 85)	Coverage	ITS-eligible
Zhihu	2,827	81 (36 pre / 45 post)	95.3%	**Yes**
Douyin	5,946	70 (25 pre / 45 post)	82.4%	**Yes**
Bilibili	5,202	16 (8 pre / 8 post)	18.8%	No
Xiaohongshu	149	13 (0 pre / 13 post)	15.3%	No
Weibo	44,626	3 (1 pre / 2 post)	3.5%	No

Coverage is reported against the common 85-month study window (2019-07 to 2026-07), not against each platform's own observation span — using per-platform spans would itself reproduce the coverage-pooling confound this study warns against. Both eligible platforms clear the 80% threshold on this common-window basis; the three ineligible platforms fall far below it even before the pre/post segment test.

Two temporal frames, not to be conflated. The eligibility counts above use the *common* 85-month window (Table 3 column "real months with ≥ 1 post of 85"). The regression in § 2.3 instead uses each platform's *own observed span* [first, last posting month] — which is why Zhihu enters with 81 months and Douyin with 79, not 85. The two frames agree on the non-empty counts (Zhihu 81/85; Douyin 70/85); the regression span additionally includes *within-span* zero-volume months (Douyin 70 non-empty + 9 empty = 79). The common-window coverage is a property of the eligibility *rule*; the per-platform span is a property of the *regression sample*. Both are disclosed so the reader can reconcile "70 months" (Table 3) with "79 months" (ITS).

Weibo warrants special comment: 44,228 of its 44,626 posts (99.1%) fall in a single month, 2022-11 — a one-time scrape snapshot, not a monthly series. Treating it as a 2019–2026 monthly series is therefore a metadata error, not a coverage gap.

2.3 ITS specification (López Bernal segmented regression)

For each eligible platform we fitted the standard segmented regression [10, 11]:

$$y_t = \beta_0 + \beta_1 \cdot \text{time} + \beta_2 \cdot \text{post} + \beta_3 \cdot \text{post_time} + \varepsilon_t$$

where time is the month index, post is an indicator for months on or after the November-2022 interruption, and post_time is the number of months since the interruption (0 before). β_2 is the immediate level change; β_3 is the slope change. The interruption was set at 2022-11-01.

Monthly-series construction rule (stated). The volume series for each platform is built on that platform's own observed span [first, last posting month] — Zhihu 2019-11–2026-07 (81 months), Douyin 2020-01–2026-07 (79 months). Three cases of a "zero" must be kept apart. (i) *Pre-platform months* — calendar

months before the platform's first observed post (e.g., 2019-07 to 2019-10 for Zhihu) — fall outside the observed span and are excluded from the regression sample; the common-window coverage reported in § 2.2 is a separate eligibility property, not part of this series. (ii) *Within-span zero-volume months* — months inside the observed span that returned zero retrieved posts — are coded as y = 0 and **retained, not dropped**; only Douyin has any (9 of its 79 months; Zhihu has none, 81 of 81 non-empty). (iii) *Sparse months* (non-zero but very few posts) are retained as observed and raise a precision concern, not a definitional one. We verified that the fill-0 construction is immaterial to the conclusion under two separate checks. First, re-fitting on the *common* 85-month window (2019-07–2026-07), padding the pre-platform months with zeros, yields near-identical coefficients — Zhihu level 63.6 (p_HAC = .072) and slope − 2.77 (p_HAC = .028); Douyin slope 4.64 (p_HAC = .037). Second, listwise-dropping the within-span zero months instead of zero-filling them yields negligibly different coefficients — Zhihu level + 65.54 (p_HAC = .064), slope − 2.62 (p_HAC = .039); Douyin slope + 4.78 (p_HAC = .032), identical to two decimal places to the fill-0 estimates — because Zhihu has no within-span zeros (dropping is a no-op); Douyin's 9 zero months are not literally leverageless as OLS data points, but they contribute no information beyond y = 0, so their deletion changes the estimates only negligibly. Both checks leave the FDR conclusion unchanged: the common-window padding check gives smallest q = .056 and the drop-zero check gives smallest q = .064, both at m = 2 (primary family), both > .05. The "nothing survives" result therefore does not depend on whether zero months are filled or dropped, nor on the exact window endpoints.

The $post_time$ coding is not a cosmetic choice, and must be stated. Because post and post_time are collinear by construction, re-centering post_time re-parameterizes β_2 while leaving β_3 numerically invariant. On the identical Zhihu series we obtain a level change of **+ 65.5** when post_time is zeroed at the interruption month (the López Bernal convention), **+ 68.2** when it is indexed from 1, and **+ 159.9** when the untransformed product post × time is used — with $\beta_3 = -2.621$ in all three cases. Only the first estimates the quantity of interest. The third measures the vertical gap between the two fitted lines extrapolated back to time = 0, a point outside the post-period support entirely; reported as an "immediate level change" it inflates the effect 2.4-fold and, in our data, converts p = .005 into p = .0001. We therefore recommend that ITS reports state the centering convention explicitly and, where feasible, report β_2 under the zeroed coding.

HAC settings must likewise be stated. Newey–West estimates depend on the lag truncation and the reference distribution, and in short monthly series the dependence is not negligible: for the Zhihu level coefficient we obtain p_HAC = .045 (lag 1, normal reference), .054 (lag 1, small-sample corrected), and .064 (lag 3, t reference). All three lie on the same side of any conventional threshold once FDR is applied, but the spread is large enough that an unreported setting makes a borderline result unreproducible. **We adopt lag 3 with the small-sample (t) reference as the primary HAC specification throughout this paper** — it is also the specification implemented in the reproducible analysis script (its_by_platform.py) — and we report the lag-1 alternatives wherever a coefficient is near α.

Reporting requirement derived here: an ITS result is only interpretable if the monthly post count is disclosed, because an interruption that coincides with a change in the retrieval yield of the query list is not identifiable from a volume series alone — and in a topic-restricted corpus the two cannot even be separately

measured. We therefore report monthly sample sizes for every eligible platform (Section 3.1/3.3) and treat any platform whose breakpoint-month sample is near zero as uninterpretable.

2.4 Multiple-comparison correction and robustness

Benjamini–Hochberg FDR [12] is reported over a primary family of m = 2: the volume level and slope of Zhihu, the only platform that is both coverage-eligible (≥ 80% non-zero monthly coverage, ≥ 12 months each side of the interruption) and inferentially identified. Douyin clears the coverage bar but its breakpoint-density selection (layer 4) leaves its coefficients non-identifiable, so it is kept out of the primary family and reported separately as not estimable (Section 3.3). Sentiment likewise stays out of the primary family: it enters no inferential conclusion (Section 2.5) and is phase-entangled. An m = 8 sensitivity analysis adds the sentiment analogues and the two Douyin coefficients. Bonferroni and Holm corrections are also reported. For robustness we computed the direct Douyin–Zhihu slope contrast and recorded its estimand, and we place placebo interruption points (2021-11, 2023-11) and Fourier seasonal terms in the Supplementary Material.

Two implementation cautions, both of which changed our own conclusions. First, the Benjamini–Hochberg procedure is a *step-up* algorithm: adjusted values must be accumulated from the largest raw P value downward, enforcing monotonicity as $q_{(i)} = \min(q_{(i+1)}, p_{(i)} \cdot m/i)$. An earlier version of our own pipeline accumulated from the smallest P value upward; that implementation is not conservative but anti-conservative, and on our data it reported a raw P of .97 as q = .04. Every FDR figure in this paper was recomputed after the defect was found, and the correction reversed our substantive conclusion — the Zhihu coefficients we had initially described as surviving do not survive. We report this explicitly because a silent implementation error in a correction procedure is undetectable from a results table.

Second, **the family composition is itself a researcher degree of freedom**, and we now report it without the earlier version's inconsistency. Under HAC standard errors (lag-3 + t) nothing in the primary m = 2 family (Zhihu level and slope) survives BH correction (smallest q = .064; lag-1 + t: 0.054). For transparency we report two counterfactual families whose composition would flip the conclusion: a family restricted to the two *slope* coefficients (Zhihu + Douyin) yields a surviving slope (smallest q = .039, lag-3 + t) — but the Douyin member is non-identifiable under layer 4, so we do not rely on it; and the m = 8 family adding the sentiment analogues and Douyin coefficients gives smallest q = .157. We take the m = 2 Zhihu-only family as primary precisely because it contains only coefficients that are both coverage-eligible and identified; the conclusion it supports — that **no coefficient is attributable to the policy interruption** — is the one we stand behind. Investigators who report a small family without a registered analysis plan should be aware that the choice of *m* and of which coefficients compose it can decide significance as surely as the choice of α. Note that even the nominally marginal Zhihu coefficients are vocabulary-driven: decomposition shows 69% of the headline level and 39% of its slope are carried by period-biased vocabulary, and their vocabulary-controlled (retained) components do not survive (retained slope p_HAC = 0.10).

Period-biased retrieval vocabulary — a definition, and a benchmark we cannot compute. We name the fourth artifact as follows. Let n_{k,p} be the number of posts returned by retrieval term k in period p (pre/post), and let r_k = n_{k,post} / n_{k,pre} be its per-period yield ratio. A retrieval instrument is *period-*

biased when, holding the keyword *set* fixed, the r_k are systematically dispersed: some terms return orders of magnitude more content on one side of the breakpoint because the events they name post-date the interruption, not because public attention changed.

The natural diagnostic would compare each r_k against ρ, the platform's own retrieval-density ratio — the factor by which its *total* monthly output changes across the breakpoint, independently of any query. **ρ is not estimable in this corpus, and we state that plainly.** Our corpus is topic-restricted by construction: every one of the 58,750 records is bound to one of 17 COVID-related query terms (Section 2.1), so no query-free platform total exists against which r_k could be normalised. Every within-corpus surrogate we examined is either circular or contaminated: the retained subset's own aggregate ratio (1.43) is definitionally equal to the quantity it would be used to certify; the full 17-term aggregate ratio (2.87) is 52% composed of the very terms under test; the full-17 monthly-mean estimator gives 2.29 and the monthly-median estimator 1.37, while the retained-subset monthly-median gives 0.55 — a 5.2-fold range across equally 'reasonable' estimators that reverses the direction of the conclusion. A benchmark whose value is fixed by an undisclosed estimator choice is a researcher degree of freedom, not a benchmark (cf. the family-size discussion above).

We therefore do **not** formalise the diagnostic as "r_k versus ρ". We report instead the *descriptive dispersion* of r_k within the query list, which requires no external benchmark and is what any investigator can compute from a per-term yield log. In our Zhihu data that dispersion is stark and bimodal: ten terms return 706 posts pre against 1,007 post (aggregate 1.43; per-term r_k from 0.56 to 6.32), while seven terms return 25 against 1,089 (aggregate 43.6; per-term r_k from 22.0 to 185.0, with one term undefined at n_pre = 0). No threshold is needed to see the gap: the two groups do not overlap.

The operational rule we apply is likewise descriptive, and we name it exactly: a **minimum per-period cell-size rule** retaining terms with ≥ 8 posts in *each* period. It is a precision rule, not a yield-balancing rule — it does not equalise r_k, and on Zhihu it selects the ten low-r_k terms only because that platform's yield distribution happens to be bimodal (any cut between 8 and 25 selects the identical ten). Section 3.3 shows that on Douyin the same rule retains a term with r_k = 42.6, which is precisely why the two concepts must be kept apart.

Decomposition and placebo-date calibration. Two further procedures support the retrieval-vocabulary analysis. First, because monthly volume is additive across term subsets and every subset is fitted on an identical design matrix, the level and slope estimates decompose exactly into a retained-term and an excluded-term component; the excluded-term regression is therefore a direct test of the difference between the full- and retained-vocabulary estimates, avoiding the invalid inference of comparing their significance levels [13]. The decomposition is specific to the identity link on the level scale, and the component variances are not additive. Second, because the excluded-term series is near-degenerate before the interruption, we calibrate each series against an in-time placebo distribution: the same HAC t statistic is recomputed at every candidate break date leaving at least 12 months (sensitivity: 18 months) on either side, and the true transition date is ranked within that distribution [14]. We report the rank explicitly, together with the resolution floor implied by the number of candidate dates. We attach no robustness claim to slope

coefficients under this calibration, because the |t| and |β| versions of the placebo statistic disagree for slopes — the post-transition trend regressor is rescaled by each candidate break date, so slope placebo distributions are not comparable across dates.

2.5 Sentiment as a measurement caveat

The sentiment field in the analysis-ready dataset is not a pure classifier output. The labeling pipeline (paper_analysis.py) runs SnowNLP but then applies an undisclosed lexicon override: posts matching a 25-term negative lexicon (NEG_LEX) are forced to ≤ 0.35, and those matching a positive lexicon (POS_LEX) to ≥ 0.75, using np.minimum/np.maximum (lines 107–118). The override touches 27.5% of posts. Locked into the two extreme intervals sit 89.1% of all sentiment values, and the neutral class shrinks to 5.4% (a rule-based proxy standard would put it near 48%; no human double-coding was done). Another 11.5% of values (6,760 posts) exceed 1.0, reaching 1.25, which contradicts the project's repeated claim of a 0–1 scale. The lexicons are also thematically entangled with the treatment variable. NEG_LEX holds the Chinese terms "death / post-COVID syndrome / medicine stockpiling / medicine panic-buying / Lianhua (a herbal antiviral)", all characteristic of the post-relaxation infection wave. POS_LEX holds "science / order / cooperation / stability", the fixed phrasing of pre-relaxation official communication. So the sentiment variable is in large part a phase indicator, and any "emotion unchanged across the pivot" null is a property of the instrument rather than evidence of public emotional stability.

Classifier agreement with an independent lexicon was Cohen κ = **0.1198 (n = 1,000)**; a rule-proxy confusion-matrix recomputation gave κ = 0.154 (n = 778). Per-platform κ ranged from 0.072 (Zhihu) to 0.224 (Xiaohongshu). These values fall far below the 0.6–0.8 range expected of validated Chinese-language sentiment instruments [15]. Because agreement is low, **no human double-coding was performed**, and the variable is phase-entangled, sentiment enters no inferential conclusion and is reported only as a cautionary measurement example.

Why corpus cleaning cannot rescue the field. A natural remedy is to remove information-type posts (testing-agency advertisements, complaint letters) and re-run. It fails for a structural reason. The lexicon override forces 90.4% of all posts into the two locked polarity intervals (≤ 0.35 or ≥ 0.75), leaving the neutral class at only 4.6% of the full sample. Even after excluding *all* information-type posts, the residual forced-polarity rate remains 90.6% (Wilson 95% CI [86.5%, 93.6%]; 345/377 merged) — barely below the 91.5% observed across all posts — because the override is a property of the instrument, not of the text. The earlier comparison of forced-polarity rates between information- and non-information-type posts returned p = .455, but that test had only 16.7% power to detect the observed 3.0-percentage-point gap (α = .05, two-sided); p = .455 therefore permits no conclusion about a difference, and in any case cannot rescue a field whose ceiling is structural. The only valid remedy is a different measurement instrument with human-labeled validation, not re-analysis of the present scores.

2.6 P9 miscapture cleaning (transparent rules)

For the 2025–2026 window (P9, n = 6,617), we applied transparent regular-expression rules flagging off-topic game fan-fiction, testing-agency advertisements, property-owners'-committee complaint letters, and unverified variant names. Flagged posts were **labeled, not silently deleted**, and retained in the analysis-

ready file with a p9_flag column. Real-world virological anchors were probed by keyword to assess content-event alignment.

3. Results

3.1 Platform coverage map

Weibo dominates the corpus with 44,626 posts (76.0%), ahead of Douyin (5,946, 10.1%), Bilibili (5,202, 8.9%), Zhihu (2,827, 4.8%), and Xiaohongshu (149, 0.3%). The annual split is heavily concentrated: 2022 alone contributes 45,449 posts (77.4%), almost all from Weibo's November scrape. Each of 2019–2021 and 2023–2024 adds under 4%. That skew is the first hint that "volume" and "coverage" measure different things.

3.2 The inference-evolution case (core demonstration)

Methodological rigor, rather than the underlying data, dictates whether a policy effect is inferred. The same data, passed through successive disciplinary checks, yield a conclusion that evolves from a spurious "significant" effect to no identifiable interruption—as summarized in Fig. 1. Evolution of the inferred conclusion under successive disciplinary checks applied to the same data. A naive pooled interrupted time-series (ITS) analysis returns a spurious "significant" immediate level change (+ 8,225.7 posts); applying the eligibility rule, heteroskedasticity- and autocorrelation-consistent (HAC) standard errors with Benjamini–Hochberg false-discovery-rate (FDR) correction, and a within-platform retrieval-vocabulary check successively removes every estimable policy effect, leaving the interruption unidentified.

3.3 Contrast: naive versus disciplined

A naive pooled analysis suggests a large, significant policy-level effect. The disciplined version shows this is arithmetic: it comes from summing platforms whose temporal supports do not compare, and from the outsized leverage of Weibo's 2022-11 scrape. Yet enforcing coverage does not recover an interpretable result. It uncovers three more artifact layers, and once all four are in view no policy-effect claim remains estimable in this corpus.

Breakpoint-density collinearity: what it breaks and what it does not. In a retrospectively assembled corpus the monthly cell size *n_m* is a property of the retrieval instrument, not a nuisance parameter. Our Douyin series carries a median of 2 posts per month before the interruption — 104 posts spread across 34 calendar months, 9 of them structurally empty, against 5,842 afterwards — a 48.5-fold jump in the median located exactly at the cut. Two distinct failures follow, and they are not of equal standing. The first is a *measurement-error* failure: the outcome $\bar{y}$_m is a sample mean with sampling variance σ^2_w/n_m, so precision changes discontinuously at the cut. Because this error enters the **dependent** variable, it does not bias the level and slope coefficients; it renders residuals sharply heteroskedastic and invalidates conventional and even HAC standard errors, which do not accommodate a deterministic discontinuity in error scale. This is a failure of *inference*, and inverse-variance weighting repairs it. The second is a *selection* failure and is decisive: retrospective retrieval on recommendation-ranked platforms returns only content

that has remained discoverable, so the sparse pre-period is a survivorship sample of viral posts while the dense post-period is a bulk sample. Empirically, the median like count of the 104 pre-breakpoint Douyin posts is 2,632 against 4 post-breakpoint (Bilibili: 633 vs 9). The two segments do not estimate the same conditional expectation; the coefficients confound any policy response with a population replacement. No weighting scheme repairs this, because the selection term shifts at precisely the cut point. Where the outcome is post volume the collapse is total: the dependent variable *is* the crawl yield, so "adjusting for monthly sample size" is algebraically degenerate (adding n_m returns R^2=1.000).

Population swap across platforms (Simpson's paradox). The naive pooled series is not a single population over time: platform composition flips completely at the interruption — before 2022-11 the corpus is 96.1% Bilibili+Zhihu, after it is 79.8% Weibo (Weibo is only 0.2% pre-interruption). The "pre/post comparison" therefore contrasts a 2019–2022 Bilibili/Zhihu sample with a November-2022 Weibo snapshot, with near-zero common support. Topic shares reverse accordingly: pooled Topic-0 share falls 13.7 pp while rising 21.0 pp within Zhihu and 11.0 pp within Bilibili. The pooled coefficient is a Simpson's-paradox confound [16], not a within-population change. The composition flip is compounded by a query-list flip that the platform labels conceal: query-list breadth ranges from 17 terms (Zhihu) to a single term (Bilibili, all 5,202 posts), so the pooled pre-period is dominated by a one-term Bilibili series alongside a seventeen-term Zhihu series, while the post-period is dominated by a one-term Weibo snapshot. Pooling therefore aggregates not merely incomparable populations but incomparable measurement instruments.

Within-platform analysis is not a sufficient remedy. The last layer is the most instructive. Zhihu is the platform whose monthly retrieved volume looks most stable across the breakpoint (median 19 pre vs 26 post posts/month), and its volume coefficient is the one a conventional infoveillance analysis would most likely headline as a policy signal. Yet it fails for a reason no between-platform correction can reach: its retrieval vocabulary is period-biased. (Fig. 2)

That apparent stability is itself an artifact of the vocabulary under test, and we flag it rather than lean on it. The 19-versus-26 medians are computed over all 17 terms; remove the seven period-biased terms and the retained subset’s monthly median drops from 20 to 11. What looks like a stable instrument is a declining retained-vocabulary series offset by a surging period-biased one. So we make no claim that Zhihu’s retrieval density is stable; we claim only that it is the platform where the decomposition can be carried out most fully.

We stress the precise form of this failure, since a coarser claim would be wrong. The keyword set is essentially stable: 16 of Zhihu’s 17 search terms appear in both periods, and none is unique to the pre-period. What moves is the yield per term. Seven post-pandemic terms — "second infection", "tested positive", "influenza A", "influenza B", "influenza-A-or-COVID", "COVID virus", and the generic "COVID" — return 25 posts before the breakpoint and 1,089 after (43.6×). The ten terms that yield substantively in both periods return 706 versus 1,007 (1.43× in aggregate, per-term ratios 0.56 to 6.32). The two groups differ by two orders of magnitude, and no external benchmark is needed to see it. Part of the apparent post-breakpoint surge is therefore an accounting feature of the query list, not a behavioural change: a term like "second infection" could hardly retrieve pre-2022 content, so including it mechanically loads volume onto

the post-period. How large that loading is, and whether the rest is behavioural, is the question the rest of this section takes up — and, as we show, it cannot be answered in this corpus. Re-estimating on the ten terms with ≥ 8 posts per period (N = 1,713) gives a level change of + 20.35 (95% CI − 34.65 to + 75.35; p_HAC = .47) and a slope of − 1.60 (p_HAC = .10). The level interval still contains the original + 65.5, so this re-estimation is too imprecise to choose between a null and the original "significant" level; it does not remove an effect it cannot tell apart from zero. To report it as an effect that keyword balancing removed would repeat the error Gelman and Stern [13] named: reading a contrast between a significant and a non-significant estimate as a finding in itself. The rest of this subsection replaces that contrast with three analyses that do not depend on it.

Exact decomposition, not a contrast of significance levels. Monthly volume is additive across term subsets and the ITS design matrix is identical for every subset, so the level estimates decompose exactly:

β_full (+ 65.54) = β_retained (+ 20.35) + β_excluded (+ 45.20).

The identity holds sample by sample — β_full − β_retained *is* β_excluded — so fitting the same ITS to the excluded-term series constitutes a direct test of the difference between the two estimates. That difference is **+ 45.20 (95% CI + 20.17 to + 70.22; p_HAC = .001)**: 69.0% of the headline level change is carried by 7 of 17 query terms (Table 4, Table 5). Two caveats belong with the identity. It is a property of the identity link on the level scale and does not survive log, square-root or Poisson specifications, which is one reason we do not adopt them as the main specification. And the *variances* do not decompose: the two components covary (Cov ≈ 135, corr ≈ 0.38), so the standard errors cannot be added. The substantive conclusion, however, is robust to the link function: under Poisson and negative-binomial specifications the excluded-term level jump remains a large, highly significant positive jump (incidence-rate ratio ≈ 50.7, p < .001) and the full-vocabulary jump likewise stays positive and significant (IRR ≈ 3.8, p < .001), so the identity-link decomposition is a presentational convenience rather than the source of the finding. (We note the excluded-term series is near-degenerate before the interruption — 25 posts across 40 months — so the Poisson IRR ≈ 50.7 is driven almost entirely by the post-period level against a near-zero baseline and should be read as a directional link-robustness check, not a precise estimate.)

Calibration against alternative break dates. *The excluded-term series is near-degenerate before the transition (25 posts across 40 months), so its HAC asymptotics are unreliable and p_HAC = .001 should not be read at face value. We therefore calibrated the same statistic (|t| on the level coefficient) against its empirical distribution across every candidate break date leaving ≥ 12 months on each side — 57 dates, giving a descriptive resolution floor of 1/57 (the smallest attainable rank; not a significance threshold) — and repeated the exercise at ≥ 18 months (45 dates) (*Table 5*). This is an in-time placebo distribution rather than randomization inference in the Fisherian sense, since break dates are not randomly assigned and the placebo statistics are mutually dependent; what it calibrates is the damage done to the HAC approximation by the trend, autocorrelation and heteroskedasticity of each particular series at T = 81. A complementary, dependence-adjusted test corroborates the rank: a moving-block bootstrap (block $\ell \in \{4,6,8\}$, B = 2000, percentile-t intervals; full distributions S7) resamples residuals drawn from the* no-break *fit* $\hat{y} = \beta_0 + \beta_1 \cdot time$ ($\beta_2 = \beta_3 = 0$), *holding the break date fixed at the true November-2022 interruption; the statistic is studentized, t =*

$\beta_2/SE(\beta_2)$, with the standard error recomputed in each of the B = 2000 resamples (percentile-t), and the block resampling acts on the residual series to reconstruct $y = \hat{y}_{null} + e$ — this implements the null of no break and re-estimates the level jump. The bootstrap P conditions on the data-driven selection of the seven ≥ 8-post-yielding terms; it certifies a discontinuity in this selected series, not policy attribution. For the excluded-term series the bootstrap P is ≈ .006–.013 — confirming the jump is real and tightly localized, in agreement with its rank-1/57 extremeness — while for the full vocabulary it is ≈ .14–.16, consistent with the non-significant HAC p = .064. The bootstrap P and the placebo rank answer different questions: the former gauges whether the level genuinely jumps under the series' own temporal dependence; the latter whether that jump is specific to the transition date. Three findings follow. First, the full-vocabulary transition date sits mid-distribution under both rules (rank 30/57; 23/45): the headline coefficient carries no date-specific information. That makes it **uninformative, not null** — having just shown the aggregate to be underpowered, we cannot read its non-rejection as evidence of absence. Second, the excluded-term discontinuity is the most extreme of all candidate dates under either rule (rank 1/57; 1/45), and its four strongest placebo dates are 2022-11, 2022-12, 2022-10 and 2022-09, so the jump is real and tightly localized at the transition. Third — and this is what the diagnostic cannot do — placebo robustness certifies only that a discontinuity exists at that date; it does not adjudicate what produced it.

The robust component is the unidentifiable one.** *Partitioning the seven excluded terms by whether they were already in wide circulation before the transition separates them sharply. Four pre-existing terms ("COVID", "COVID virus", "influenza A", "influenza B") contribute + 13.34 (20.4%); for terms of this kind a zero or near-zero pre-period yield — "COVID virus" returns 0 posts before the transition and "COVID" only 7 across 40 months — is consistent with recency bias in the retrieval index, but is equally consistent with genuinely low pre-transition salience or with pre-transition discussion under different lexical items; within this corpus the three are observationally equivalent, so the recency-bias reading is a hypothesis, not a finding. Three post-transition coinages ("second infection", "tested positive", "influenza-A-or-COVID") contribute + 31.86 (48.6%), and these are exactly the terms for which the two competing explanations are indistinguishable: a public that genuinely began discussing reinfection in 2023, and a retrieval index weighted toward recent content, predict the same series. The calibration ranks the coinage component first of 57, whereas the pre-existing component reverses across trimming rules (descriptive rank 7/57 versus 1/45). The reversal has a specific cause: the ≥ 18-month rule removes a cluster of large negative placebo statistics in 2025 that reflect non-linear decay of the same 2022-11 shock, left unmodelled by the linear post-transition trend; under that rule the calibration also assigns both components the same lowest attainable descriptive rank and so loses all power to tell them apart. Every placebo-robust component of the effect therefore lies inside the unidentified region.

***What can and cannot be concluded.** *The share attributable to vocabulary-yield asymmetry is bounded only within [0%, ≥ 69%]. The lower bound is not established, because the one component that can be attributed cleanly to the instrument is not robust to the calibration window. The upper bound is not sharp either, because the retained set itself retains per-term pre/post yield ratios ranging from 0.56 to 6.32 (*Table 4*) under an inclusion threshold (≥ 8 posts) that is ultimately arbitrary. The residual retained-vocabulary estimate is distinguishable neither from zero nor from the full-vocabulary estimate whose value its interval contains. We therefore conclude that the policy-associated level change is not estimable** in this corpus — neither confirmed nor excluded — rather than absent or demonstrably artifactual.

***Why the aggregate is weaker than its own component.** *That a component is significant while the total is not is arithmetically consistent, not contradictory. Adding the retained-term series raises the level estimate by 45% (from + 45.20 to + 65.54) but raises its HAC standard error by 174% (from 12.77 to 34.94), because the retained series contributes substantial break-unrelated low-frequency variation; the signal-to-noise ratio therefore falls, and |t| drops from 3.54 to 1.88. Aggregation over a heterogeneous retrieval vocabulary is not variance-free: a well-localized discontinuity in one component can be masked in the aggregate, so the aggregate null result is not evidence against the component result. This is a general hazard for infoveillance designs that sum a query list into a single series. (Under the narrower 79-month frame spanned by the subset itself the estimates are + 21.97 and − 1.47, the same conclusion; we report the 81-month frame for consistency with the main analysis.) The threshold is not tuned: any cut-off between 8 and 25 posts per period selects the identical ten terms. We note for completeness that the* literal *intersection criterion — all 16 terms appearing in both periods — leaves the estimates essentially unchanged (+ 64.86, p_HAC = .064), because it readmits the seven period-biased terms on the strength of one to seven pre-period posts each. This contrast is itself the methodological point: "keywords common to both periods" is too weak a standard; common* yield, *not common* presence, *is what must be held constant. The finding is further undermined by a single month (2022-12 contributes 442 posts, 21.1% of Zhihu's post-period; excluding it drops the level to + 28.82, p_HAC = .076) and by leave-one-out instability (level significant in only 1 of 81 refits). Under HAC standard errors with correct Benjamini–Hochberg step-up across the m = 8 family (which for completeness also includes the non-identifiable Douyin coefficients and the four sentiment analogues), the smallest q is 0.157 — no coefficient is attributable to the policy interruption. This m = 8 family is a liberal sensitivity check; because the Douyin coefficients are non-identifiable (*Section 2.3*), they are not valid tests, and the m = 8 result must not be read as conferring validity on them. Finally, placebo breakpoints confirm the diagnosis for Douyin: pseudo-interruptions in 2020-11, 2021-05, 2021-11 and 2022-05 all return* larger* volume slopes (+ 5.04 to + 6.04, p ≤ .003) than the true breakpoint (+ 4.78, p = .032), whereas Zhihu's volume placebos are uniformly null (p = .072–.914).

Table 4

Per-term retrieval yield for Zhihu's 17 query terms, by period. Note. Terms are grouped by the ≥ 8-posts-per-period inclusion rule; the ten retained terms retain ratios spanning 0.56–6.32, which is why the 69% upper bound is not sharp. The 'Retained subtotal' and 'Excluded subtotal' ratios (1.43 and 43.6) are aggregate yield ratios *of the respective selected subsets*; neither is an external instrument benchmark, and 1.43 must not be read as the platform's crawl-density ratio (see Section 2.4). 'COVID virus' returns 0 posts pre-breakpoint, so its r_k is undefined rather than large.

Group	Query term	Pre	Post	Post/Pre
Retained (10)	nucleic-acid testing	121	68	0.56
Retained	COVID vaccine	111	71	0.64
Retained	fever clinic	115	74	0.64
Retained	Omicron	103	83	0.81
Retained	COVID pneumonia	48	94	1.96
Retained	epidemic prevention and control	44	90	2.05
Retained	antigen	59	125	2.12
Retained	COVID sequelae	49	117	2.39
Retained	COVID symptoms	31	127	4.10
Retained	do we still wear masks	25	158	6.32
Retained subtotal (aggregate, not a benchmark)	**10 terms**	**706**	**1,007**	**1.43**
Excluded — pre-existing (4)	COVID	7	184	26.3
Excluded — pre-existing	COVID virus	0	68	∞
Excluded — pre-existing	influenza A	6	187	31.2
Excluded — pre-existing	influenza B	7	154	22.0
Excluded — coinage (3)	second infection	1	185	185.0
Excluded — coinage	tested positive	1	173	173.0
Excluded — coinage	influenza-A-or-COVID	3	138	46.0
Excluded subtotal	**7 terms**	**25**	**1,089**	**43.6**

Table 5

Exact decomposition of the Zhihu volume level change and in-time placebo calibration. Note. All estimates use Newey–West HAC standard errors (maxlags = 3, t-based) on Zhihu's shared 81-month frame; components sum exactly to the full-vocabulary estimate. The placebo columns give the *descriptive rank* of the true transition date among all candidate break dates — not a P value: candidate break dates are mutually dependent and un-randomised, so no look-elsewhere correction is available. A dependence-adjusted significance level from a moving-block bootstrap (S7) is reported alongside: excluded-term level jump P_boot≈ .006–.013; full vocabulary P_boot≈ .14–.16.

Series (terms)	Level	95% CI	p_HAC	Placebo rank ≥ 12 mo	Placebo rank ≥ 18 mo
Full vocabulary (17)	+ 65.54	−2.93 to + 134.02	.064	30/57	23/45
Retained (10)	+ 20.35	−34.65 to + 75.35	.47	54/57	42/45
Excluded, all (7)	**+45.20**	**+20.17 to + 70.22**	**.001**	**1/57**	**1/45**
pre-existing (4)	+ 13.34	+ 1.57 to + 25.11	.029	7/57	1/45
coinages (3)	+ 31.86	+ 12.52 to + 51.19	.002	1/57	1/45
Retained + coinages (13)	+ 52.20	−13.16 to + 117.56	.12	48/57	—

***The same decomposition on Douyin.** *The vocabulary artifact is not a Zhihu peculiarity. Douyin's query list comprises nine terms, four of which return zero pre-breakpoint posts, and 87.0% of its corpus (5,170 of 5,946) comes from three post-transition coinages ("second infection", 1 vs 2,679; "tested positive", 35 vs 1,490; "influenza B", 0 vs 965); the list contains no pre-relaxation vocabulary at all — no "nucleic-acid testing", "vaccine", "Omicron", "epidemic prevention and control" or "antigen". Applying the same ≥ 8-posts-per-period rule retains three terms, and the slope change decomposes exactly as + 4.78 (p_HAC = .032) = + 0.34 (three retained terms; 95% CI − 2.15 to 2.83; p_HAC = .79) + 4.44 (six excluded terms; 95% CI 2.08 to 6.80; p_HAC = .0004)*: 92.9% of the Douyin slope is carried by terms that could barely retrieve pre-breakpoint content (Table 6). Two points follow. Unlike Zhihu, the retained-vocabulary estimate here is not merely imprecise — it is centred essentially at zero (t = 0.27), so for Douyin the vocabulary explanation is not merely admissible but is consistent with accounting for essentially the whole coefficient (treated descriptively, given Douyin's non-identifiability). And the retained set includes "tested positive" at r_k = 42.6, which demonstrates the point made in Section 2.4: a minimum-cell-size rule is a precision rule, not a yield-balancing rule, and the two coincide on Zhihu only because that platform's yield distribution happens to be bimodal.

Table 6

Exact decomposition of the Douyin volume slope change. Note. Components sum exactly to the full-vocabulary slope. The sub-coefficients are reported *descriptively* to show how the slope partitions; they are not independent hypothesis tests, and they are *not* subjected to Benjamini–Hochberg correction. Douyin is excluded from the primary m = 2 FDR family (Section 2.4) because its breakpoint-density selection makes its coefficients non-identifiable, so no family correction is applied to its decomposition.

Series (terms)	Slope	95% CI	p_HAC
Full vocabulary (9)	+ 4.78	0.48 to 9.09	.032
Retained (3)	+ 0.34	−2.15 to 2.83	.79
Excluded (6)	**+4.44**	**2.08 to 6.80**	**.0004**

3.4 P9 content quality and event alignment

The 2025–2026 window (n = 6,617; Bilibili 3,200, Douyin 2,548, Weibo 392, Zhihu 341, Xiaohongshu 136) was screened with transparent rules that flagged 42 posts (0.6%): 19 testing-agency advertisements, 19 property-owners'-committee complaint letters, 3 game fan-fiction items, and 1 post claiming an unverified "mute-strain" variant. Keyword probes for four real-world virological anchors turned up only 2 posts mentioning XEC and 1 mentioning LP.8, while NB.1 (26 posts) and chikungunya (5 posts) appeared almost entirely in 2026 although their real epidemiological windows fall in 2024–2025. This segment is therefore both sparse and misaligned in time with the events it should track: a transparency case study, not evidence of full seven-year evolution.

4. Discussion

4.0 The identification limit of retrospective retrieval

The central methodological result is negative, and it generalizes — but only in a structural sense we make explicit here. The quantitative findings do not transport. The 69% vocabulary share, the Douyin slope decomposition, and the near-degenerate Weibo snapshot all depend on Chinese platform quirks and a single 2022-11 breakpoint; none should be read as estimates for other languages, platforms, or health topics. What does generalize is the identification problem itself. A zero or near-zero pre-period yield for a query term arises from three mechanisms at once: the phenomenon had not happened yet, the lexical item was not in circulation, or the retrieval index is skewed toward recent content (a classic instrumentation threat in quasi-experimental design [17]; the instability of search-engine trend instruments is also well documented [18]). In a retrospectively assembled corpus without a retrieval vocabulary frozen before collection, these mechanisms are observationally equivalent [17], and no within-corpus contrast can tell them apart. Our data make the point in its sharpest form: the only component of the Zhihu level change that survives placebo-date calibration is the one contributed by terms that entered Chinese after the interruption, exactly the component for which instrument change and genuine topic emergence predict the same series. External evidence on when a term first entered circulation (search-index first-appearance data,

say) could in principle bound the problem, but that needs re-collection rather than re-analysis, and we make no such claim. The limit is plain: absent a pre-registered, frozen and version-controlled retrieval vocabulary, an interrupted time series on platform search output cannot separate a change in the retrieval instrument from a change in the public it is meant to measure. This is not peculiar to our corpus; it is structural to retrospective social-media infoveillance, which is why we place vocabulary freezing among the prospective-collection requirements rather than the reporting suggestions. Importantly, this identification-limit claim stands independent of whether the Zhihu coefficients happen to survive FDR: even a statistically significant $q < .05$ would not rescue policy attribution, because the limit is logical (observational equivalence), not statistical.

One further observation is worth recording, since it bears on how much weight such diagnostics can carry. Our placebo calibration is itself specification-dependent: the pre-existing-term component sits at descriptive rank 7/57 under the 12-month rule but rank 1/45 under the 18-month rule. A diagnostic designed to detect specification-dependent artifacts thus inherits, rather than eliminates, analytic degrees of freedom. We report both rules side by side for that reason.

4.1 Reporting norms implied by this study

Six write-up standards, plus two prospective-collection requirements set out below, should become standard for social-media infoveillance. They build on Jandoc's ITS reporting review [19], the TREND guideline for non-randomized evaluations [20], and a recent methodological review calling for better conduct and reporting of interrupted time-series analyses [21], and extend that lens into the retrieval-vocabulary domain. Because ITS is non-randomised, the analysis also falls under the ROBINS-I bias domains [22], and the Section 1.4 taxonomy maps each artifact onto one such domain. The family-composition and lag choices that decide significance (Section 2.4) are covered by standard 3 and must be pre-specified rather than chosen post hoc:

1. 1. **Timestamp provenance disclosure** — state explicitly, per platform, whether the temporal field is a publication date or a scrape-batch timestamp, and disclose *both*. In our analysis CSV the field pub_time is a verified publication date for all platforms, but the crawl-batch timestamp crawl_time was not joined; we report this gap rather than assume publication dates are absent.
2. 2. **Eligibility rule with monthly-coverage disclosure** — pre-specify (or transparently post-specify) a coverage threshold and report each platform's real month count, not merely the span between first and last observation.
3. 3. **Multiple-comparison correction with pre-specified family composition and lag** — report FDR with the family size *m* **stated and pre-specified**, including exactly which coefficients compose the family and why (a coefficient judged non-identifiable must be excluded from the primary family, not averaged in); state the HAC lag-selection rule and report all alternatives near α (we report lag-1 and lag-3). Pair FDR with leave-one-out stability and placebo breakpoints as defaults, not bare P-values.

4. 4. **post_time coding declaration** — state the centering convention for the post_time regressor; in our data re-coding it alone moves the reported level change from + 65.5 to + 159.9 on an identical series, so the convention must be fixed and reported, not left implicit.
5. 5. **Platform-composition disclosure per period** — report the platform mix before and after any interruption; a pooled pre/post contrast must not silently be a population swap (Simpson's paradox). Where common support is near zero, pooled inference should be abandoned (see Fig. 2).
6. 6. **Retrieval-instrument stability disclosure** — report the retrieval keyword list together with its **per-period yield** (posts returned per term per period), report the *dispersion* of the per-term yield ratios r_k rather than a single aggregate, and re-estimate on the subset of terms that clear a stated minimum per-period cell size — naming that rule for what it is (a precision rule), since it does not by itself equalise yield. We note, as an instance of this standard applied to ourselves, that the quantity one would most want here — the platform's query-free retrieval-density ratio ρ, against which r_k could be normalised — **cannot be computed from a topic-restricted corpus such as ours, and we could not compute it.** Disclosing the benchmark one cannot compute is part of the standard, not an exemption from it. Investigators who can obtain a query-free platform volume series — a platform-side sampling frame, or a deliberately off-topic control query crawled on the same schedule — should report r_k/ρ; those who cannot should say so explicitly. Verifying that a term merely *appears* in both periods is insufficient: in our data that weaker criterion left the estimate intact (+ 64.86, p_HAC = .064) whereas equalizing yield left it statistically indistinguishable from the unadjusted level (level + 20.35, 95% CI − 34.65 to + 75.35, containing + 65.5; p_HAC = .47) — the confidence interval is too wide to adjudicate between a null and the original estimate. Query lists assembled after the event tend to contain period-specific vocabulary that can retrieve content from only one side of the breakpoint, mechanically manufacturing a level shift. Within-platform analysis is *not* a sufficient remedy for composition confounding: in our data 69% of the headline coefficient proved to be carried by seven of seventeen query terms, and the part of it that withstands placebo-date calibration is precisely the part that cannot be attributed to the instrument rather than to the public. Report the decomposition and its bounds; do not report the retained re-estimate as an effect that has been *removed*. Report HAC (or otherwise heteroskedasticity-robust) standard errors, leave-one-out stability, and placebo breakpoints alongside the main estimate; note that a placebo breakpoint returning a *stronger* effect than the true one, as ours did for Douyin, falsifies the design rather than the hypothesis.

Prospective collection (requirements 7–8). Standards 1–6 govern write-up; two further requirements govern collection that is a *necessary precondition* for breakpoint inference — they narrow, but do not close, the identification limit set out in Section 4.0; full identification still requires an external anchor (e.g., search-index first-appearance data), which demands re-collection rather than re-analysis: (7) **freeze and version-control the retrieval keyword list before crawling** and log per-term monthly yield so retrieval-instrument stability is auditable; (8) **store** $crawl_time$ **and** $publish_time$ **in separate fields** and ensure ≥ 12 months of comparable coverage on each side of any planned interruption. Teams without a per-term yield log should not attempt breakpoint inference. (The four collection actions (a)–(d) listed in the Abstract are consolidated here into requirements 7–8.)

4.2 Content moderation as a competing mechanism

A volume jump at a policy interruption is not, on its own, proof that the public reacted. On Chinese platforms, moderation (deletion, down-ranking, hot-search intervention) is selective and far from random [23, 24]. A "no structural sentiment break" result is exactly what selective moderation produces, and it breaks the ITS assumption that measurement stays constant. So a null or weak result in this setting may be a moderation artifact rather than evidence that nothing happened. That is why every finding here is framed as associational and platform-specific.

4.3 Limitations

Limitations separate into a universal structural layer — retrieval-instrument change vs genuine topic emergence are observationally equivalent without a frozen vocabulary; pooled series incur Simpson's-paradox population swaps; sentiment instruments entail phase entanglement — and a study-specific layer (one 2022-11 Chinese policy pivot, a specific moderation ecology, Chinese lexical specifics) that does not transfer to other languages, platforms, or topics (cf. Section 4.0).

1. Weibo's dates sit in a single 2022 scrape window (99.1% in 2022): a snapshot, not a monthly series; the 2022-11 concentration cannot be temporally disaggregated. (2) Verified pub_time exists for every platform, but crawl_time was never joined; we disclose this gap rather than treat dates as missing. (3) Sentiment is a lexicon-override artifact (27.5% overwritten; 11.5% >1.0), phase-entangled with the policy ($\kappa = 0.12$) — a phase indicator, not a validated emotion measure. (4) The eligibility rule is post hoc (threshold sensitivity provided). (5) On eligible platforms the interruption coincides with a yield change: Douyin's monthly count jumps 48.5× at the breakpoint (104 posts over 34 pre months, 9 empty, vs 5,842 after), a survivorship sample (median likes 2,632 pre vs 4 post). Adding monthly sample size to a volume ITS is degenerate ($R^2=1.000$); inverse-variance weighting mends sentiment-ITS SEs but not the estimand. (6) External validity is doubly bounded: the identification failure generalizes, but the estimates hinge on one policy transition and a Chinese ecology. (7) The pooled series swaps populations at the interruption (Bilibili/Zhihu-led pre, Weibo-led post; common support ≈ 0) — a Simpson's paradox, not a within-population change; report composition per period, not only volume. (8) Within-platform analysis does not fix (7): Zhihu's vocabulary is period-biased — seven terms return 25 vs 1,089 posts and carry + 45.20 of + 65.54 (69.0%), leaving a residual of + 20.35 (95% CI − 34.65 to + 75.35; p_HAC=.47), ≈ 0. The set is stable (16/17 terms both periods), and co-presence does not dissolve the effect (+ 64.86; p_HAC=.064); per-term yield is operative. This is a failure of identification, not proof of no effect: the placebo-robust remainder comes from post-transition coinages, where instrument change and topic emergence are observationally equivalent. This study reports no surviving policy effect; all coefficients are artifact demonstrations, not claims about Chinese public discourse.

5. Conclusions

Multi-platform social-media ITS is valid by temporal completeness, timestamp provenance, and a stable-yield retrieval instrument — not corpus size. Enforcing temporal integrity collapsed our "five-platform, seven-year" story to two platforms and then undid that: under HAC errors with FDR step-up, no primary-family

coefficient survived (smallest q=.064 for the m = 2 Zhihu level/slope family; .157 at m = 8), none attributable to the interruption (independent of the marginal FDR result).

The largest residual coefficient is not a policy effect: 69% rides on seven of seventeen terms (transition rank 30/57), and on Douyin 92.9% of the slope rides on six of nine terms — four with no pre-breakpoint content — while the placebo-robust remainder is post-transition vocabulary, where instrument change and genuine discussion are observationally equivalent.

We propose as standard for social-media infoveillance: scrape-batch timestamps, eligibility rules with coverage disclosure, monthly-sample-size and platform-composition reporting, per-period keyword disclosure, placebo breakpoints, and FDR correction. The cautionary case has four artifact layers under two analysis layers (Section 1.4): coverage pooling (composition flip — Simpson's paradox, near-zero support), sentiment phase entanglement (measurement failure), period-biased vocabulary (our principal claim), and breakpoint-density selection (Douyin's estimand/selection failure, decisive alone). The sharpest lesson: a single platform — the usual remedy for platform confounding — is insufficient, because the data-generating instrument can shift at the breakpoint, indistinguishable from the behaviour it measures without a frozen, version-controlled query list.

Declarations

Data Availability Statement

The corpus is 58,750 public posts from five Chinese platforms (Weibo, Douyin, Bilibili, Zhihu, Xiaohongshu) via 17 COVID-19 terms. **Raw posts cannot be redistributed** under platform terms of service (single-investigator exemption, not a data-sharing release); de-identified exports — the per-term/month/platform volume matrix, per-term per-period yield tables, and P9 miscapture flag — are available to qualified researchers on reasonable request, subject to those terms.

Code (Newey-West HAC ITS, level/slope decomposition, breakpoint placebo calibration, moving-block bootstrap; Python 3.10, statsmodels 0.14) is **Supplementary File S9**, to be deposited in a public repository (GitHub/Zenodo) with a citation DOI; reproducible scripts and CSVs are open-licensed. The Douyin illustration uses only derived monthly counts. This secondary analysis of public, non-identifiable data needed no new ethics approval and was not prospectively registered.

Ethics approval and consent to participate: This study was approved by the Research Ethics Committee of Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology (approval number TJ-IRB20230846). Because only publicly posted, non-sensitive social-media content was analyzed and all data were de-identified, the requirement for individual informed consent was waived. All methods were carried out in accordance with relevant guidelines and regulations.

Clinical trial number: not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The analysis-ready, de-identified data and code supporting this study are described in the Data Availability Statement and provided as Supplementary Material S1, S8 (data tables) and S9 (analysis code). Restrictions apply to the raw retrieved posts under the platforms' terms of service; derived aggregate tables are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: None declared.

Authors' contributions: S. Q. and Y. C. contributed equally to conceptualization, data curation, formal analysis, and manuscript drafting. L. X. conceived the study, supervised the analysis, and finalized the manuscript. All authors read and approved the final version.

Acknowledgements: Not applicable.

References

1. World Health Organization. Coronavirus disease 2019 (COVID-19): situation report, 13. 2 February 2020 [cited 2026 Aug 14]. Available from: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf
2. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res. 2009;11(1):e11. 10.2196/jmir.1157.
3. Tangcharoensathien V, Calleja N, Nguyen T, et al. Framework for Managing the COVID-19 Infodemic: Methods and Results of an Online, Crowdsourced WHO Technical Consultation. J Med Internet Res. 2020;22(6):e19659. 10.2196/19659.
4. White B, Ishizumi A, Yau B, et al. Narrative Trends over the COVID-19 Pandemic: Digital Social Listening to Inform WHO Infodemic Management. Stud Health Technol Inf. 2023;302:893–4. 10.3233/SHTI230297.
5. Pilipiec P, Samsten I, Bota A. Surveillance of communicable diseases using social media: A systematic review. PLoS ONE. 2023;18(2):e0282101. 10.1371/journal.pone.0282101.
6. Lazer D, Kennedy R, King G, et al. The parable of Google Flu: traps in big data analysis. Science. 2014;343(6176):1203–5. 10.1126/science.1248506.
7. Ma X. Research on social media information security questions in the big data era.Advances in Computer Science Research, 2016. 10.2991/iccsae-15.2016.78
8. Olteanu A, Castillo C, Diaz F, et al. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Front Big Data. 2019;2:13. 10.3389/fdata.2019.00013.
9. Langan SM, Schmidt SA, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ. 2018;363:k3532. 10.1136/bmj.k3532.

10. Wagner AK, Soumerai SB, Zhang F, et al. Segmented regression analysis of interrupted time series studies in medication use research. J Clin Pharm Ther. 2002;27(4):299–309. 10.1046/j.1365-2710.2002.00430.x.
11. López Bernal J, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol. 2017;46(1):348–55. 10.1093/ije/dyw098.
12. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57(1):289–300. 10.1111/j.2517-6161.1995.tb02031.x.
13. Gelman A, Stern H. The difference between significant and not significant is not itself statistically significant. Am Stat. 2006;60(4):328–31. 10.1198/000313006X152649.
14. MacKinnon JG, Webb MD. Randomization inference for difference-in-differences with few treated clusters. J Econom. 2020;218(2):435–50. 10.1016/j.jeconom.2020.04.024.
15. Zhao N, Jiao D, Bai ST, et al. Evaluating the validity of Simplified Chinese version of LIWC in detecting psychological expressions in short texts on social network services. PLoS ONE. 2016;11(6):e0157947. 10.1371/journal.pone.0157947.
16. Hernán MA, Clayton D, Keiding N. The Simpson's paradox unraveled. Int J Epidemiol. 2011;40(3):780–5. 10.1093/ije/dyr041.
17. Stone-Romero EF. Experimental and quasi-experimental designs for generalized causal inference. Organ Res Methods. 2004;7(2):224–7. 10.1177/109442810472006.
18. Nuti SV, Wayda B, Ranasinghe I, et al. The use of google trends in health care research: a systematic review. PLoS ONE. 2014;9(10):e109583. 10.1371/journal.pone.0109583.
19. Jandoc R, Burden AM, Mamdani M, et al. Interrupted time series analysis in drug utilization research is increasing: systematic review and recommendations. J Clin Epidemiol. 2015;68(8):950–6. 10.1016/j.jclinepi.2014.12.018.
20. Des Jarlais DC, Lyles C, Crepaz N, TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. Am J Public Health. 2004;94(3):361–6. 10.2105/ajph.94.3.361.
21. Korevaar E, Karahalios A, Turner SL, et al. Methodological systematic review recommends improvements to conduct and reporting when meta-analyzing interrupted time series studies. J Clin Epidemiol. 2022;145:55–69. 10.1016/j.jclinepi.2022.01.010.
22. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919. 10.1136/bmj.i4919.
23. King G, Pan J, Roberts M. How Censorship in China Allows Government Criticism But Silences Collective Expression. Am Polit Sci Rev. 2013;107(2):326–43. 10.1017/S0003055413000014.
24. King G, Pan J, Roberts ME. Political science. Reverse-engineering censorship in China: randomized experimentation and participant observation. Science. 2014;345(6199):1251722. 10.1126/science.1251722.

Figures

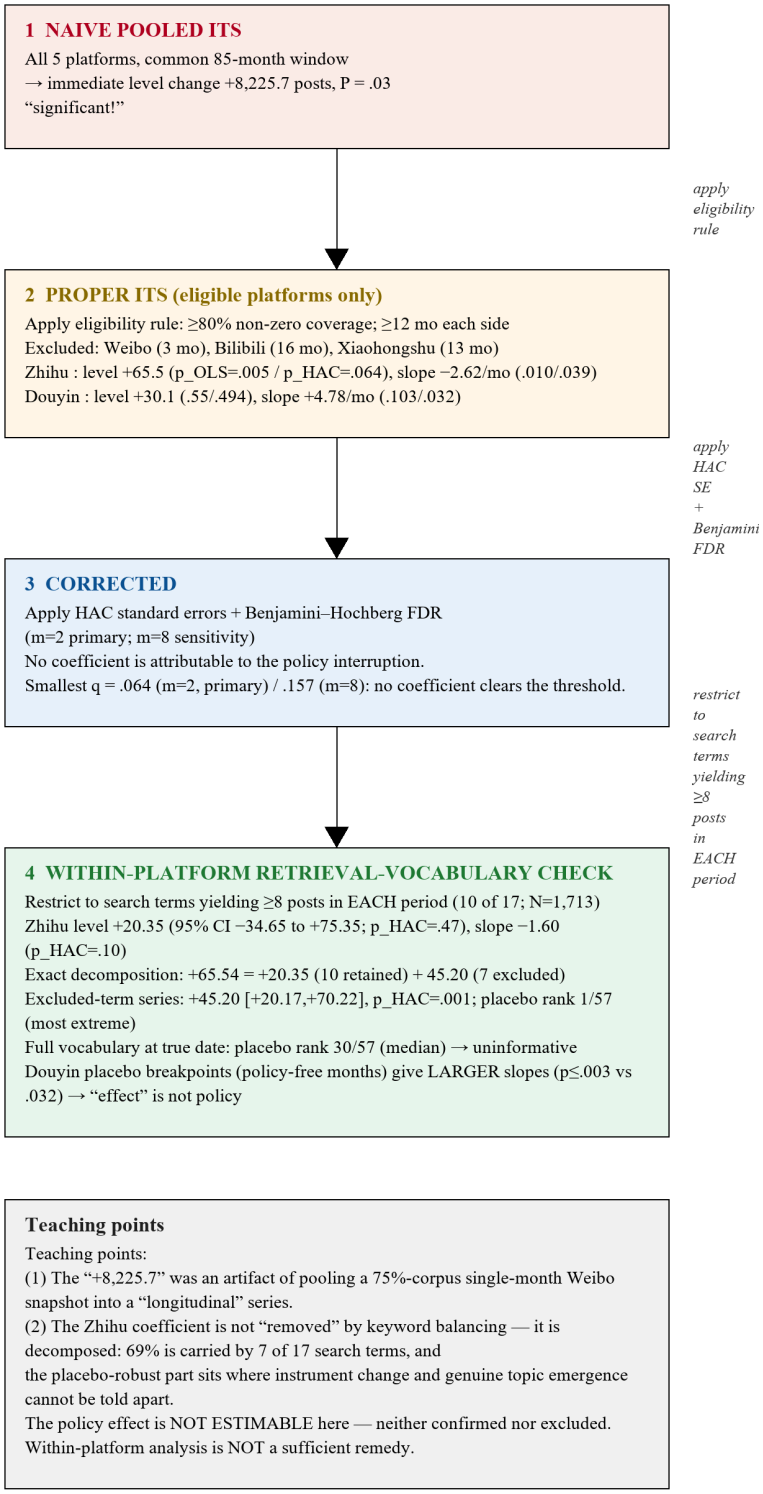


Figure 1

Evolution of the inferred conclusion under successive disciplinary checks

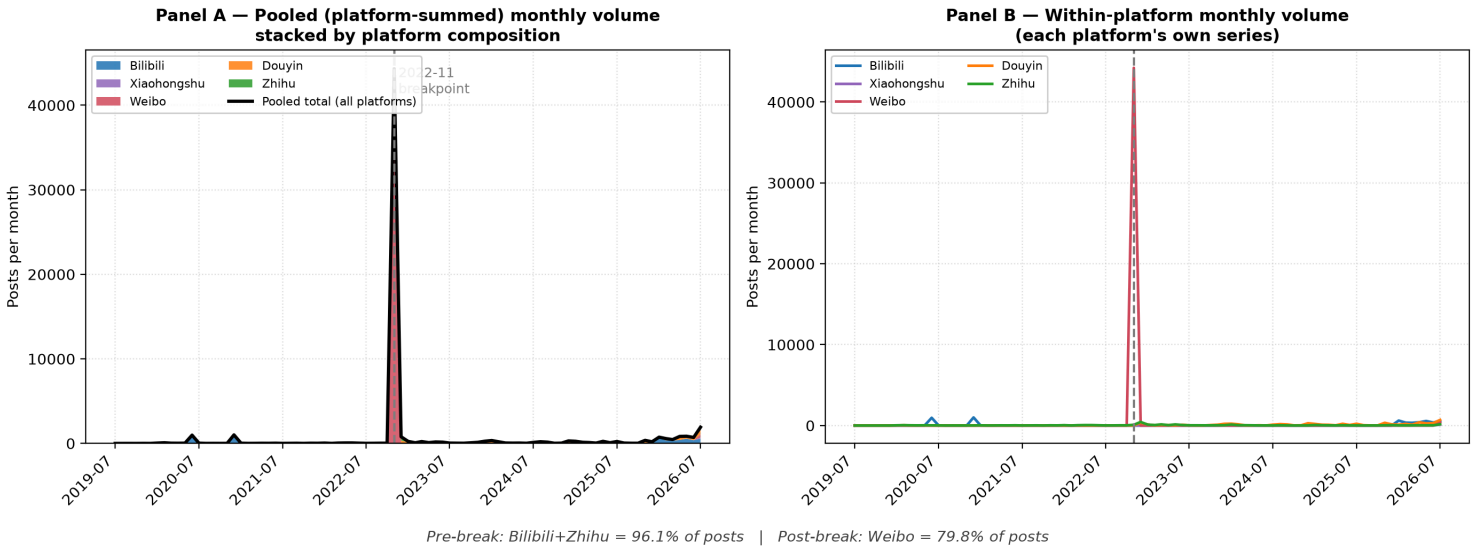


Figure 2

Pooling masks the breakpoint platform-composition reversal. Panel A plots the monthly total volume across all five platforms (black line) with each platform's contribution stacked. Pre-breakpoint, Bilibili+Zhihu jointly account for 96.1% of posts; post-breakpoint, Weibo alone accounts for 79.8% — the stacked-area colour composition flips accordingly. Panel B plots each platform's own within-platform monthly volume. The dramatic pooled jump at 2022-11 is entirely attributable to a single Weibo scrape batch concentrated in that month; no other platform shows a comparable shift. The pooled coefficient is therefore a Simpson's-paradox confound between two near-disjoint populations, not a within-population change. The figure is the empirical face of §4.1#4.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryCombinedS1S10.docx
- SupplementaryMaterialinfo.docx