What the Corpus Still Holds

Testing keys, survival and result stability in deposited YouTube comment corpora

Max Sedlmair · ORCID 0009-0007-2471-5595 · Independent Research, Germany · research@woidwide.net

Version 1.0 · 30 October 2026 · Not peer reviewed. Refutations, review, or comments are much appreciated.
Data collected 26–30 October 2026 (YouTube Data API v3).
This paper: https://doi.org/10.5281/zenodo.23805575 · Verification package (scripts, outputs, design documents): https://doi.org/10.5281/zenodo.23805596
Part of a series on YouTube comment corpora: What the Corpus Does Not Say, https://doi.org/10.5281/zenodo.23339873 · What the Corpus Can No Longer Find, https://doi.org/10.5281/zenodo.23642703

Abstract. Researchers often share YouTube comment data as comment IDs plus their own labels and expect others to retrieve the comments through the YouTube Data API. Whether this works, and what is lost, has rarely been measured on real deposits. This study is meant for anyone who builds, shares or reuses such data: linguists, discourse analysts and communication researchers in particular.

I tested three public deposits in October 2026, from the oldest to the youngest:

- 62,414 comments on COVID-19 videos, collected in 2020, almost half of them labelled hostile;
- 173,012 comments on the U.S. Army's channel, collected in 2023;
- 16,652 comments on precision fermentation, collected in 2024.

Each step was fixed before it was run. Every deposited figure was first reproduced from the file. A sample of IDs was looked up one by one to confirm that the keys still work. Then every comment was checked, and the deposited results were recomputed on what remained.

Findings:

- The keys still work.
- 74.4%, 93.2% and 94.8% of the comments, from oldest to youngest deposit, can still be retrieved.
- The loss comes in blocks: whole videos and whole threads. On YouTube, a reply almost never outlives the comment it answers.
- The more hostile the corpus, the more its hostile comments are lost. The difference is 4.8 percentage points in the oldest and most hostile corpus (5.4 on videos that still exist), slight in the contested one, and absent in the low-conflict one.
- Yet every deposited distribution moves by less than one percentage point.
- A re-collection by listing alone misses comments wherever IDs are damaged or videos are not listed. In one deposit, the video ID of 1,845 comments has been replaced by `#NAME?`, the error value a spreadsheet shows when it reads text as a formula.

One consequence: a reply that corrects a claim disappears with the claim.

For distributions, a loss of this size is a nuisance rather than a threat. For analyses of threads or of individual users it may be a threat. The paper closes with five checks for anyone who shares or reuses YouTube comment data. These are three deposits measured once; the results are offered as a starting point, not a general law, and comments, corrections and refutations are welcome.

The paper is the third part of a series on what YouTube comment corpora can and cannot support. The three papers complement each other: the first looks at what a corpus leaves out when it is collected, the second at whether a deposited corpus can still be traced back to the platform, and this one at how much of it survives and whether the loss changes its results.

Keywords: YouTube; comment IDs; rehydration; data decay; data persistence; reproducibility; hate speech; corpus linguistics; computer-mediated communication; YouTube Data API

1. Introduction

Researchers who share YouTube comment data increasingly share only the comment IDs together with their own labels, and invite others to retrieve the comments through the YouTube Data API. Two recent deposits say so in as many words (Chan et al., 2026; Cruickshank et al., 2024). The practice rests on two assumptions:

- the IDs still work as keys years later;
- what can no longer be retrieved does not change the result.

Both depend on a condition outside the researchers' control: that the platform keeps its API open to research (Davidson et al., 2023).

The first assumption failed in an earlier case. The comment IDs of a 2017 research deposit no longer resolve at all, although most of the comments they pointed to are still visible on the platform under different IDs (Sedlmair, 2026b). That study could not ask the second question, because no key worked.

This study asks it for deposits whose keys still work. It measures how much of three deposited YouTube comment corpora is still observable two to six years after collection, and where the loss falls. It asks whether patterns emerge in which comments disappear: by thread position, by age, and by the stance or hostility the deposit assigned to them. Finally, it asks whether the loss changes the quantities the deposits report.

That content disappears from platforms is well known, and so is the fact that the loss is not random:

- for Twitter, older datasets lose a growing share of their posts (Zubiaga, 2018);
- deletions cluster in particular accounts and political camps (Bastos, 2021; Elmas, 2023);
- in shared abusive-language datasets, harmful items are lost at different rates from others, and some items become available again (Assenmacher et al., 2023);
- recovering a deposited corpus and recomputing its published results can change them materially (Küpfer, 2024);
- in reused datasets, retrieval loss is often label-imbalanced, with hate speech and misinformation labels losing more (Ansari et al., 2023);
- on Twitter, replies can remain accessible after the post they answer is deleted (Mondal et al., 2016). Whether the same holds for YouTube has not been shown.

For YouTube:

- comments on 258 political videos were collected in January 2018 and re-collected in June 2018. Comments that had disappeared were more frequent on ideologically extreme and false videos and after a fact-check (Jiang, Robertson & Wilson, 2019);
- comments on Italian COVID-19 videos were re-collected at least one year after posting, and 32% of the violent comments were no longer available (Cinelli et al., 2021);

- a five-day observation of 407,463 comments on 500 recent videos recorded 9,738 removals and 263 edits (Wang & Weller, 2025, pp. 36–39). The authors call for longer observation periods.

A related line of work concerns corrections:

- false rumours corrected by a fact-check comment were deleted more often (Friggeri et al., 2014);
- posts with visible Community Notes are withdrawn more often (Gao et al., 2026; Chuai et al., 2026);
- on Reddit, posts with a fact-check comment were removed by their authors less often but by moderators more often (Bond & Garrett, 2023).

This study does not claim that attrition, or its selectivity, is new. It is a specific empirical extension of established persistence and replication research. What it adds is a measurement on real YouTube deposits, in a fixed order: check that the keys work, measure what has become unavailable and where, then recompute the deposited results.

The study is the third part of a small research programme on YouTube comment corpora:

- *What the Corpus Does Not Say* (Sedlmair, 2026a) asks what a corpus leaves out at collection.
- *What the Corpus Can No Longer Find* (Sedlmair, 2026b) asks whether deposited keys still lead back to the platform.
- This study asks, once the keys work, how much survives and whether the loss matters.

The three studies complement each other. Each covers one stage in the life of a corpus: collection, deposit and the years after. Each also relies on the others: a survival test needs keys that work, and keys are only useful if what was collected is understood.

The programme has a practical purpose. Linguists, discourse analysts and communication researchers increasingly build their findings on YouTube comments, and they need to know which of those findings will hold up and remain checkable years later. Each study therefore ends with concrete checks, so that results drawn from platform data are not only correct on the day of collection but remain reproducible afterwards.

All three studies follow one rule: test a premise before relying on it, fix the test in writing before running it, and report a failed premise as a result.

2. The deposits

Three deposits were tested. Each is public, licensed for reuse, contains YouTube comment IDs in the current format (`Ug…`), and carries per-comment labels assigned by its authors. None contains text that this study reads.

Table 1. The three deposits.

	Chan et al. (2026)	DIVERSE (Cruickshank et al., 2024)	IMSyPP English (Ljubešić et al., 2021)
Topic	precision fermentation	the U.S. Army's YouTube channel	COVID-19 videos
Collected	2024	2–5 October 2023	February–May 2020
Comments (of which replies)	16,652 (4,354)	173,012 (51,204)	62,414 (18,078)
Videos	180	859	174
Labels used here	sentiment, GPT-4o and EmoRoBERTa	stance toward the Army and the video; sentiment; hate speech; sarcasm (model labels)	type of hate speech, two human annotators per comment
Deposited result recomputed	disagreement between models, 39.1%	stance shares (Table III, arXiv v2)	annotation distribution (project report D2.2)
Licence	CC BY 4.0	CC BY 4.0	CC BY-SA 4.0

Step A (no API). Before any query, every deposited figure that the study recomputes was reproduced from the deposit. The reproduction was exact in every case:

- **Chan:** the 39.11% disagreement (6,512 of 16,652) and all six label shares.
- **DIVERSE:** the six stance counts of Table III. The percentages printed there do not match the counts and sum to about 95%; the counts were used.
- **IMSyPP:** the four annotation counts of the train set (52,990 / 1,739 / 45,863 / 2,589).

For DIVERSE, the paper reports 173,013 comments on 858 videos. The deposit contains 173,012 comments under 859 values in the video column, two of which are not valid video IDs: `#NAME?`, a spreadsheet formula error (1,845 comments), and a 13-character string (108 comments). All comment IDs are well formed. This study uses the deposit as it is (Addendum A11).

DIVERSE and IMSyPP contain comment texts and author fields. Those columns were never loaded.

3. Method

3.1 Order and pre-specification

Each step was fixed in writing before it was run, in a freeze and its addenda (Section 8). For every deposit the order was the same:

1. reproduce the deposited figures from the file (Step A);
2. test whether the IDs are valid keys (the gate);
3. measure which comments are observable;
4. describe the loss (D1, Section 4), and freeze that description;
5. recompute the deposited results on the observable comments only, once (D2, Section 5).

No analysis was added after D2.

The deposits were added one after another, not chosen together.

- Chan was planned first.
- DIVERSE was added before its data were opened, to give a contested topic.
- IMSyPP was added after DIVERSE showed a small tendency against hostile comments, to test that tendency where hostile content is common.

Each addition was fixed in an addendum before any query on the new deposit. Readers should know that the third deposit was selected because of what the second suggested; the test itself was fixed before its data were queried.

3.2 Are the keys valid?

An earlier study found that IDs from a 2017 deposit no longer resolve at all (Sedlmair, 2026b). Each collection therefore began with a validity gate:

- 100 deposited IDs (50 top-level, 50 replies, fixed seed) were looked up individually with `comments.list`;
- if none was found, the run would stop.

The gate found 94 (Chan), 95 (DIVERSE) and 66 (IMSyPP) of 100. In every deposit the keys were valid.

3.3 Which comments are observable?

Listing. For each available video, all current top-level threads were listed (`commentThreads.list`, `part=id`). For each parent of a deposited reply, all current replies were listed (`comments.list`, `parentId`). A deposited ID counts as present if it appears in these listings. Only IDs were requested; no text.

Cross-check against the deposits' own rehydration route. A random sample of listing-absent IDs was looked up one by one:

- 300 for Chan, 1,000 for DIVERSE and IMSyPP;
- plus 200 present IDs in each deposit.

The rule, fixed in advance: if more than 2% of the absent IDs are found by individual lookup, the listing is incomplete, and all absent IDs are looked up individually.

Individual lookups are needed because `comments.list` answers a batch of IDs with HTTP 404 if any one of them is missing (Sedlmair, 2026b).

Table 2. Gate, cross-check and final count.

	Chan	DIVERSE	IMSyPP
Gate: IDs found	94 / 100	95 / 100	66 / 100
Present in listing	15,649	159,347	46,458
Cross-check: absent IDs found individually	36 / 300 (12%)	132 / 1,000 (13%)	1 / 1,000 (0.1%)
Full individual lookup of absent IDs	yes (1,003)	yes (13,665)	no (rule not met)
Found only by individual lookup	133	1,897	–
Observable	**15,782**	**161,244**	**46,458**
Collected	26 October 2026	26–29 October 2026	29–30 October 2026

Observable means returned by the API on the collection date. It does not mean visible to the public, and it says nothing about who removed a comment or why (Section 6).

3.4 What was stored

- one presence bit per deposited ID;
- per video, its availability and today's comment count;
- the gate and cross-check counts.

No text, no author and no other field was retrieved or stored. All analytical variables come from the deposits.

3.5 Analysis

D1 (survival) describes the observable share overall, by video status, top-level vs reply, by comment year, and by every deposited label.

Uncertainty. All intervals are 95% cluster-bootstrap intervals over videos (2,000 draws, fixed seeds), because comments under one video are not independent.

Differential rule, fixed in advance. Loss counts as differential if two categories of the same variable differ by more than 2 pp and the interval of the difference excludes 0.

Conditional analysis (IMSyPP only, pre-specified). The main contrast was repeated on the videos still available. This separates the removal of whole videos from the removal of individual comments.

D2 (result survival) recomputes each deposited result on the observable comments only, once. Δ = survivors − full deposit, in pp, with a cluster-bootstrap interval. The pre-specified reading:

- **practically unchanged:** |Δ| ≤ 1.0 pp;
- **changed:** |Δ| > 1.0 pp and the interval excludes 0.

Primary quantities:

- Chan: the disagreement share;
- DIVERSE: the share “against” toward the Army;
- IMSyPP: the share of hostile annotations (offensive or violent) in the train set.

4. Results I: what is still there

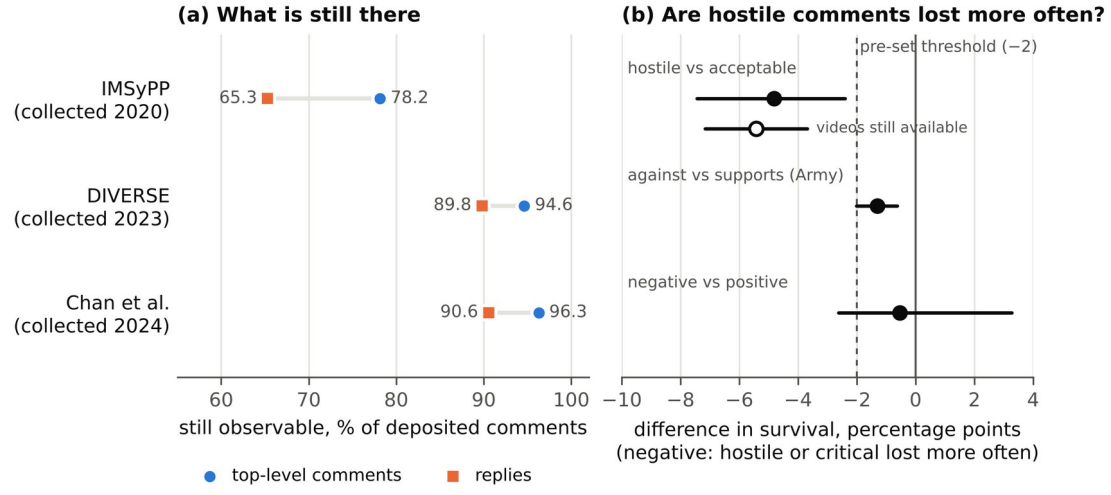
4.1 How much can still be retrieved

In October 2026, between three quarters and nineteen twentieths of the deposited comments could still be retrieved (Table 3, Figure 1a). The two youngest deposits had lost least. The oldest, which is also the most hostile, had lost most.

Table 3. Comments still observable, by deposit.

	Chan (2019–2024)	DIVERSE (≤ 2023)	IMSyPP (2020)
Still observable	**94.78%** [92.11, 96.37]	**93.20%** [92.52, 93.77]	**74.44%** [69.54, 78.64]
Videos no longer available	8 of 180	9 of 857, plus 2 malformed IDs	15 of 174
Videos with no comment left	10	12	15
Median share left per video	100%	95.7%	86.3%

Figure 1. (a) Share of deposited comments still observable, for top-level comments and replies. (b) Difference in survival between hostile or critical comments and acceptable or supportive ones, with 95% intervals. The dashed line is the threshold fixed before the analysis (−2 percentage points). The open circle shows IMSyPP restricted to videos that are still available.



4.2 Where the loss falls: whole videos and whole threads

Most of the loss comes in blocks, not one comment at a time.

Whole videos.

- In all three deposits, no comment is left on a video that is no longer available.
- In DIVERSE, two values in the video column are not valid video IDs: `#NAME?` (1,845 comments) and a 13-character string (108 comments). These comments can still be looked up by their own ID, and 92% of them are still there. They are counted, but not as comments on unavailable videos (Addendum A11).

Whole threads. When a top-level comment disappears, its replies almost always disappear with it (Table 4). This is why replies are lost more often than top-level comments in all three deposits.

Table 4. Replies and the comments they answer.

	Chan	DIVERSE	IMSyPP
Top-level comments still observable	96.27%	94.63%	78.16%
Replies still observable	90.56%	89.79%	65.30%
Difference	+5.71 pp [+2.56, +8.96]	+4.84 pp [+3.70, +5.95]	+12.86 pp [+9.27, +15.74]
Above the pre-set threshold?	yes	yes	yes
Replies whose parent is gone, still observable	6 of 249 (2.4%)	66 of 2,941 (2.2%)	**0 of 3,221**

4.3 By age

Within the two younger deposits, **newer** comments were lost slightly more often than older ones:

- Chan: comments from 2024 against 2019, +4.56 pp [+2.13, +7.32];
- DIVERSE: comments from 2023 against 2022, +3.08 pp [+1.96, +4.56].

Both differences are above the threshold. Across deposits, the oldest one (IMSyPP) has lost the most. That comparison, however, mixes age with topic and hostility (Section 7).

4.4 By label: are hostile or critical comments lost more often?

Table 5. Survival by the label the deposit assigned.

Deposit	Comparison	Difference in pp [95% interval]	Above the threshold?
Chan	negative vs positive sentiment (GPT-4o)	−0.53 [−2.62, +3.27]	no
Chan	the two models agree vs disagree	−0.01 [−1.45, +0.75]	no
DIVERSE	against vs supports (the Army)	−1.31 [−2.01, −0.62]	no (smaller than 2 pp)
DIVERSE	against vs supports (the video)	−1.47 [−2.14, −0.80]	no (smaller than 2 pp)
DIVERSE	hate vs no hate	−1.79 [−3.13, −0.50]	no (smaller than 2 pp)
DIVERSE	negative vs positive sentiment	−1.05 [−1.98, −0.02]	no
IMSyPP	**hostile vs acceptable (all videos)**	**−4.82 [−7.45, −2.40]**	**yes**
IMSyPP	**hostile vs acceptable (videos still available)**	**−5.42 [−7.16, −3.68]**	**yes**
IMSyPP	violent vs acceptable (videos still available)	−1.71 [−4.14, +0.22]	no

The pattern rises in steps (Figure 1b):

- **Chan:** no sign that negative comments are lost more often.
- **DIVERSE:** critical, negative and hateful comments are lost slightly more often. Every comparison points the same way and none of the intervals includes zero, but every difference stays below the 2-point threshold.

- **IMSyPP:** here almost half the comments are hostile, and the difference is clearly above the threshold. It remains when only videos that still exist are counted, so it is not explained by whole videos disappearing.
- **Violent comments:** comments that any annotator marked as violent are not lost more often than acceptable ones.

In short: across these three deposits, the more hostile the corpus, the more often its hostile comments are lost. The deposits also differ in age and topic, so this ordering describes; it does not explain (Section 7).

---

5. Results II: do the deposited results still hold?

Table 6. Deposited figures, recomputed on the comments that are still observable.

Deposit	Figure (main figure in bold)	Full deposit	Still observable	Change in pp [95% interval]	Reading
Chan	**disagreement between the two models**	39.11	39.11	+0.00 [−0.18, +0.38]	practically unchanged
Chan	largest change among the six label shares			≤ 0.07	practically unchanged
DIVERSE	**against, toward the Army**	36.80	36.61	−0.20 [−0.33, −0.06]	practically unchanged
DIVERSE	supports, toward the video	32.54	32.89	+0.35 [+0.18, +0.52]	practically unchanged
IMSyPP	**hostile annotations, train set**	46.96	46.32	−0.64 [−1.36, +0.20]	practically unchanged
IMSyPP	offensive annotations	44.45	43.75	−0.70 [−1.37, +0.09]	practically unchanged

“Practically unchanged” means a change of at most one percentage point, as fixed in advance.

Every deposited distribution moves by less than one percentage point. The direction matches Section 4: critical and hostile shares shrink slightly, and supportive or acceptable shares grow slightly.

Further checks.

- In Chan, agreement with the human coder changes by −0.16 pp (GPT-4o) and −0.22 pp (EmoRoBERTa).
- In DIVERSE, the most supportive and the most neutral video stay the same. The three videos with the highest “against” share lie within 0.9 pp of each other and change order.

6. Discussion

The keys work, and most of the corpus is still there. For deposits with current IDs, rehydration by ID works. After two to three years, about 5–7% of comments are unavailable; after six and a half years, in a COVID-19 corpus, about a quarter. This is the counterpart to the earlier study, where no key worked (Sedlmair, 2026b).

Loss is organised by videos and threads. Most unavailable comments belong to a few unavailable videos or to threads whose top-level comment is gone. On YouTube, a reply almost never outlives its parent (0.0–2.4%). On Twitter, replies can survive deletion of the post they answer (Mondal et al., 2016). For corpus users this means:

- reply-level and thread-level data decay faster than top-level data;
- the loss comes in blocks, not at random.

The more hostile the corpus, the more its hostile comments are lost.

- No difference in a low-conflict corpus (Chan).
- A small, consistent tilt in a contested corpus (DIVERSE).
- A clear, differential loss in a hostile-rich corpus (IMSyPP), surviving the restriction to videos that still exist.

This fits what is known for Twitter (Elmas, 2023; Ansari et al., 2023; Assenmacher et al., 2023). It extends the one-year YouTube observation of Cinelli et al. (2021): there, 32% of violent comments were unavailable after at least one year; here, 25% of violent comments are unavailable after about six and a half years, in the English sister corpus of their Italian collection (74.77% still observable, Section 4).

The API does not tell who removed a comment: the author, the channel owner or the platform. No cause is attributed.

Yet the deposited distributions hold. In all three deposits, every reported share moves by less than one percentage point, including the hostile share in IMSyPP. The reason is arithmetic. When most of the loss falls on all categories alike, a few points of extra loss in one category shift its share only slightly.

This reassurance is limited to distributions. Analyses that depend on the lost comments themselves are more exposed: threads, reply chains, individual users, or the most hostile videos. For distributions, attrition of this size is a nuisance, not a threat; for thread- and user-level analyses it may be a threat.

A listing finds only what it can reach. In two deposits the pre-specified cross-check found many comments that the listing had missed: 12% in Chan and 13% in DIVERSE. Almost all of them lay where a listing cannot reach:

- Chan: 127 of 133 were on unlisted videos, which the collector did not list;
- DIVERSE: 1,797 of 1,897 were under two malformed video IDs in the deposit, one of them the spreadsheet error `#NAME?` (Addendum A11).

On intact IDs of listed videos, the listing was almost complete: 6 comments in Chan, 0.9% of the listing-absent comments in DIVERSE, and 1 of 1,000 in the IMSyPP cross-check. The full individual lookup captured all of these comments, so the observable shares are correct.

Two practical points follow:

- a listing-only re-collection overstates loss wherever IDs are damaged or videos are not listed;
- deposited ID columns can be damaged when they pass through a spreadsheet: an ID that begins with `-` may be read as a formula and replaced by an error value.

The cascade also takes corrections with it. On YouTube, a reply that disputes a claim is a reply like any other. When the claim is deleted, the correction goes too; in IMSyPP, none of 3,221 such replies survived. A corpus re-collected later therefore lacks the claim and also lacks the record that it was contested. This study did not classify which replies were corrections, so the size of this effect is not measured.

What researchers can do. These checks continue those of the two earlier papers in the series (Sedlmair, 2026a, 2026b):

1. Check the ID columns for format errors, then test a sample of deposited IDs one by one before rehydrating. The earlier study shows why.
2. Re-collect by listing **and** by individual lookup, or at least cross-check a sample.
3. Report survival by thread position and by label, not only overall.
4. Recompute the deposited result on the survivors, and report the difference.

5. Deposit the analytical labels with the IDs, as all three deposits here did. Without them none of this could be checked.

How far this goes. These are three deposits, measured once, on one platform. The results are offered as a starting point for checking deposited YouTube data, not as a general law of comment loss. Readers who find an error, a better explanation or a study that already answers these questions are asked to say so. Comments, corrections and refutations are welcome.

7. Limitations

- **Three deposits, chosen in sequence.** The third was chosen after the second suggested a tendency (Section 3.1). The tests were fixed before each query, but the set of deposits is not a random sample.
- **Age, topic and hostility are confounded across deposits.** IMSyPP is the oldest, the most hostile and about COVID-19, a topic with heavy platform moderation. The graded pattern across deposits cannot separate these factors. Only the within-deposit contrasts are clean.
- **Labels differ in kind:** model labels in Chan and DIVERSE, human labels in IMSyPP. The IMSyPP train set was sampled from videos with many hateful comments.
- **One measurement point** per deposit. A second collection is planned for May 2027 with the same scripts.
- **Observability is not visibility.** The API may return comments the public cannot see, and the reverse is possible. Causes of removal are not identified.
- **Edits** are not detected. A comment that survives may have changed since it was labelled.
- **Clusters in DIVERSE.** The 1,845 comments under `#NAME?` probably belong to several videos but enter the bootstrap as one cluster. The intervals for DIVERSE are therefore approximate.
- **Collection setup.** The collection ran on the author's phone (Termux): one day for Chan, four days for DIVERSE and two days for IMSyPP. One run was interrupted by the operating system and resumed from its log; no step was repeated or lost.

8. Provenance

A note from the author. My interest is primarily linguistic: I want to know what YouTube comment data can tell us about language and conversation. That question kept leading back to the data themselves, and so this series has become, I hope, useful measurement work as well. I have worked as carefully as I could, fixed every step in advance, and published the corrections I found along the way. Even so, I may have made mistakes or overlooked work that should have been cited, and I apologise to any author whose study is missing here. I would be grateful to anyone who checks the figures against the deposits and the verification package, and who tells me where I am wrong.

Each step was fixed in writing before it was run:

Document	Content
Freeze P2 v0.1	Chan: question, gate, listing, cross-check, D1/D2, decision lines
Addendum A1	DIVERSE added; repeat measurement May 2027
A2–A4	Chan: listing incomplete (pre-specified fallback), D1, D2
A5–A7	DIVERSE: Step A and collection design, D1, D2
A8–A10	IMSyPP added and designed, D1, D2
A11	Findings while building the verification package: malformed video IDs in DIVERSE, unlisted videos in Chan; offline reproduction byte-identical
Freeze H1′ v0.1 + feasibility check	Corrections and claims: not run, power too low (MDE ≥ 6 pp); kept for a later study

Document	Content
Literature checks	29 October 2026: two external searches for prior work (Consensus and a second language model), used to place the study in Section 1

AI assistance. The author designed the study, set all analytic decisions and decision rules, and revised all text. Claude (Anthropic) was used under the author's direction to draft text and analysis code, which the author then edited. All numbers come from scripts run on the deposited data. The author is responsible for the content.

9. Data and code availability

All scripts, the presence bits, the ID files and the design documents are deposited in a verification package (https://doi.org/10.5281/zenodo.23805596). The README maps every table and the figure to its script and output; every offline step reproduces byte for byte from the original deposits. The package contains no comment text and no author information.

References

Ansari, M., Sidhpura, J., Mandal, V. K., & KhudaBukhsh, A. R. (2023). Quantifying the transience of social web datasets. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2023)*. ACM. https://doi.org/10.1145/3625007.3627596

Assenmacher, D., Sen, I., Fröhling, L., & Wagner, C. (2023). The end of the rehydration era: The problem of sharing harmful Twitter research data. In *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media (NEATCLasS)*. https://doi.org/10.36190/2023.56

Bastos, M. T. (2021). This account doesn't exist: Tweet decay and the politics of deletion in the Brexit debate. *American Behavioral Scientist, 65*(5), 757–773. https://doi.org/10.1177/0002764221989772

Bond, R. M., & Garrett, R. K. (2023). Engagement with fact-checked posts on Reddit. *PNAS Nexus, 2*, pgad018. https://doi.org/10.1093/pnasnexus/pgad018

Chan, J., Wang, Y.-L., Kok-Shun, B. V., & Woo, M.-W. (2026). *A dual-annotated YouTube comment corpus on precision fermentation technology (2019–2024)* (Version 2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.21146055

Chuai, Y., Pilarski, M., Renault, T., Restrepo-Amariles, D., Troussel-Clément, A., Lenzini, G., & Pröllochs, N. (2026). Community-based fact-checking reduces the spread of misleading posts on X (formerly Twitter). *Nature Communications, 17*. https://doi.org/10.1038/s41467-026-72597-0

Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Kralj Novak, P., & Zollo, F. (2021). Dynamics of online hate and misinformation. *Scientific Reports, 11*. https://doi.org/10.1038/s41598-021-01487-w

Cruickshank, I. J., Ng, L. H. X., & Soofi, A. (2024). DIVERSE: A dataset of YouTube video comment stances with a data programming model. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE. https://doi.org/10.1109/BigData62323.2024.10825480. Preprint version 2: arXiv:2403.03334v2. Data set: Zenodo, https://doi.org/10.5281/zenodo.10493803

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., van der Linden, D., Roscoe, J. F., Ayravainen, L., & Cork, A. G. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour, 7*(12), 2054–2057. https://doi.org/10.1038/s41562-023-01750-2

Elmas, T. (2023). The impact of data persistence bias on social media studies. In *Proceedings of the 15th ACM Web Science Conference 2023*. ACM. https://doi.org/10.1145/3578503.3583630

Friggeri, A., Adamic, L. A., Eckles, D., & Cheng, J. (2014). Rumor cascades. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1). https://doi.org/10.1609/icwsm.v8i1.14559

Gao, Y., Zhang, M. M., & Rui, H. (2026). Can crowdchecking curb misinformation? Evidence from Community Notes. *Information Systems Research*. Advance online publication. https://doi.org/10.1287/isre.2024.1609

Jiang, S., Robertson, R. E., & Wilson, C. (2019). Bias misperceived: The role of partisanship and misinformation in YouTube comment moderation. *Proceedings of the International AAAI Conference on Web and Social Media, 13*(1), 278–289. https://doi.org/10.1609/icwsm.v13i01.3229

Küpfer, A. (2024). Nonrandom tweet mortality and data access restrictions: Compromising the replication of sensitive Twitter studies. *Political Analysis, 32*(4), 493–506. https://doi.org/10.1017/pan.2024.7

Ljubešić, N., Mozetič, I., Cinelli, M., & Kralj Novak, P. (2021). *English YouTube hate speech corpus* [Data set]. Jožef Stefan Institute, CLARIN.SI. http://hdl.handle.net/11356/1454

Mondal, M., Messias, J., Ghosh, S., Gummadi, K. P., & Kate, A. (2016). Forgetting in social media: Understanding and controlling longitudinal exposure of socially shared data. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association.

Sedlmair, M. (2026a). *What the corpus does not say: Reply structure, time and coverage in YouTube comment data* [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.23339873

Sedlmair, M. (2026b). *What the corpus can no longer find: Legacy YouTube comment IDs in a 2017 research deposit no longer resolve, although most of the comments are still there* [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.23642703

Wang, Y., & Weller, K. (2025). Tracking ephemerality in YouTube comments: Towards methods for building dynamic CMC corpora. In A. Fábián & I. Trost (Eds.), *Impulses and approaches to computer-mediated communication: Proceedings of the 12th International Conference on Computer Mediated Communication and Social Media Corpora for the Humanities* (pp. 36–39). University of Bayreuth.

Zubiaga, A. (2018). A longitudinal assessment of the persistence of Twitter datasets. *Journal of the Association for Information Science and Technology, 69*(8), 974–984. https://doi.org/10.1002/asi.24026