CURIOSITYBITS LIVING LAB BOSTON · STUDY #07 ACADEMIC RESEARCH ARTICLE (READER-ACCESSIBLE EDITION V1.1)

Clicks Without Conversation? Negative Valence, Moral-Conflict Framing, and the Structure of Public Deliberation in Digital Journalism

CuriosityBits Living Lab — Communication and Media Science Series (Study 7, Reader-Accessible Edition v1.1)

Provenance & Audit Status: Edition v1.1 preserves 100% of the pre-registered empirical findings, numbers, tables, and scope conditions of Edition v1.0—independently verified as real archival data and replicated by both the Boston Lab Forensic Self-Audit (November 19, 2026) and the CuriosityBits Living Lab Singapore External Red-Team Audit (Audit 37, DeepSeek V4.1 Flash, November 19, 2026)—while expanding plain-language methodological explanations, intuitive real-thread walk-throughs, a reader's guide, a glossary, and appendices for interdisciplinary social science and humanities scholars.

Plain-Language Executive Summary: What This Study Asks, How It Works, and What We Found

The Question: Does "Engagement" Mean a Healthy Public Conversation?

Across communication studies, political science, and media psychology, a familiar rule of thumb holds that **"if it bleeds, it leads"** and **"outrage drives engagement."** Dozens of experiments and platform studies show that when a news headline uses negative emotion (crisis, danger, loss, tragedy) or moral conflict (blaming an opponent, accusing an institution of wrongdoing, pitting two groups against each other), online audiences click more, share more, and leave more comments.

For scholars of democracy and the public sphere, however, lumping **clicks, upvotes, shares, and total comment counts** into a single bucket called "engagement" hides the most important question: **What kind of conversation actually happens inside the comment section?** When a headline draws people in through negativity or moral blame, does it spark a broad, inclusive, fact-based public discussion where many citizens exchange reasons and cite evidence? Or does it simply trigger a quick burst of initial reactions followed by a few locked-in combatants arguing back and forth in a corner?

Why Studying This Is Tricky (and How Our Two-Stage Design Solves It)

Comparing online comment threads across news stories is full of hidden traps: 1. **The Survival Trap (Selection Bias):** Most news links posted online receive zero or only a handful of comments. If a researcher only looks at "popular" threads that reached 50 or 100 comments, they are studying the rare survivors without

knowing how headlines shaped which stories survived in the first place. To solve this, we use a **two-stage hurdle design**: **Stage 1** studies a broad, representative sample of N = 3653 news submissions (including stories with zero comments) to see what gets a discussion started and what crosses the **15-comment front-page breakthrough hurdle**. **Stage 2** zooms inside N = 2114 active discussion trees (226317 individual comments posted within 48 hours) to map the geometry of who replies to whom. 2. **Tragedy Is Not the Same as Blame:** A headline about a devastating earthquake or a veterinary health crisis is *negative* (sad, alarming, threatening), yet blames no villain. Conversely, a headline about a Supreme Court dispute or a regulatory accusation is *conflictual* (two sides clashing over right and wrong). Instead of relying on crude word-counting dictionaries that confuse sadness with blame, we had **four distinct open-weight AI language models** independently read and score $5398 real headlines (seeing *only* the headline words, completely blind to the outlet, date, upvotes, or comments) on two separate 0.00-to-1.00$ scales: **Headline Negativity** and **Moral-Conflict and Blame Framing**. 3. **Separating Real Dialogue from "Trench Duels":** In digital deliberation research, when Speaker A posts a comment, Speaker B replies to A, and Speaker A replies back to B (A \to B \to A), researchers call this **dyadic reciprocity** and usually celebrate it as genuine back-and-forth dialogue. We go one step deeper: we split every A \to B \to A exchange into **first-time reciprocity between a new pair of citizens** versus **repeat ping-pong volleys within the exact same two-person pair** (A \to B \to A \to B \to A \dots).

Five Core Findings in Plain English

1. **Sparks Without a Wildfire (Stage-1 Crowd Gatekeeping):** Negative and moral-conflict headlines do spark an initial reaction—each boosts the chance that a story gets *at least one comment* by roughly +3.06 and +3.00 percentage points (off a $43.25% baseline) and attracts more passive crowd upvotes. Yet in a community-curated forum without a personalized outrage algorithm (Hacker News), that initial spark fizzles out quickly: **neither negative tone nor moral conflict helps a story cross the 15-comment threshold needed to become a sustained front-page discussion** ($p = 0.1315$ and $p = 0.6079$).

2. **Sadness Alone Does Not Shape Conversation (The Valence Null):** Once a discussion is underway (Stage 2), **pure Headline Negativity without moral blame has zero effect on any of our five measures of conversation structure** (all adjusted $p = 1.0000$). A grim headline about a natural disaster or medical hazard leaves the shape, depth, diversity, and fact-citing of the comment section completely unchanged.

3. **Moral Conflict Creates "Trench Duels," Not a Town Hall (Agonistic Dyadic Closure):** By contrast, **Moral-Conflict and Blame Framing shifts all five conversation metrics simultaneously** (all adjusted $p \le 0.0177$). At first glance, moral conflict appears to boost A \to B \to A back-and-forth reciprocity (+1.35 percentage points per standard-deviation increase, or +4.81 percentage points across the full $0-to-1 conflict scale, off a 21.17% average), increase nested replies (+1.27 percentage points), and push comment branches deeper (+0.1812 levels at the 90th percentile). **However, our exact mathematical decomposition reveals that 80.44% of that reciprocity increase (+1.09 of the +1.35 percentage points) is driven by the exact same two people trading repeat ping-pong blows**, while first-time dialogue between new pairs barely budges (+0.26 percentage points, statistically indistinguishable from zero, p =

0.2950). At the exact same time, the number of unique speakers per comment shrinks (-1.74 percentage points), the top two most vocal speakers take over a larger share of the thread (+0.74 percentage points), overall speaker diversity drops (-0.0029 Shannon entropy units), and **citations to external documentary evidence fall by -0.73 percentage points per standard deviation** (-2.59 percentage points across the full conflict scale, a notable drop off the 9.91%$ average evidence-citing rate). We call this pattern **agonistic dyadic closure**: moralized blame locks a comment section into deeper, narrower two-person trench duels with fewer facts.

4. **Propositional Contestability, Not Dark Mood, Drives Counter-Arguing:** Even **non-negative debates**—policy, legal, or scientific disagreements written in a calm, non-tragic tone ($N_i < 0.50$, $C_i \geq 0.50$)—significantly increase back-and-forth reciprocity (+2.12 percentage points, $p = 0.0302$), nested replies (+1.79 percentage points), and branch depth (+0.2694 levels). People argue back and forth when a headline presents a *contestable claim with competing sides*, not when it merely sounds bleak.

5. **Readers Repair Paywalls Together (Networked Epistemic Repair):** We hypothesized that news paywalls (such as *The New York Times* or *The Atlantic*) would starve non-subscribers of article facts, forcing them to fight over the headline alone. Instead, we found that when an article is paywalled, **commenters post $5.38 times as many archive-mirror links** (1.089% of comments on paywalled stories contain an archive link versus 0.202% on free stories; regression-adjusted jump +0.92 percentage points, $t = +5.73$, $p < 0.0001$). By sharing full-text mirrors at the top of the thread, the reading community neutralizes the paywall barrier so that downstream discussion behaves just like discussion on open-access outlets (interaction adjusted $p = 1.0000$$).

How to Read This Paper: A Quick Guide for Humanities & Social Science Readers

You do not need formal training in econometrics or computational statistics to follow every argument in this paper. Here is how the paper is organized and how to read the numbers in the tables:

- **What a "One-Standard-Deviation (z) Change" Means:** Because our two headline scores—**Headline Negativity** and **Moral-Conflict Framing**—are continuous ratings between $0.00 (none at all) and 1.00$ (intense), we report effects in two intuitive ways: 1. **Per Standard Deviation (`Neg (z)` and `Conflict (z)` in tables):** A one-standard-deviation shift (roughly +0.28 points on the $0.00-to-1.00 scale) represents a typical step up from an average headline (say, 0.29 conflict) to a clearly conflictual one (around 0.57$ conflict). 2. **Across the Full [0, 1] Scale:** Comparing a completely non-conflictual headline ($0.00) to a maximally conflictual headline (1.00) multiplies the per-standard-deviation shift by roughly 3.55. For example, Moral Conflict's +0.0135 (+1.35 percentage-point) per-SD effect on dyadic reciprocity corresponds to a **+0.0481 (+4.81 percentage-point) jump across the full [0, 1] scale**—moving a thread from the 21.17% average reciprocity rate up to roughly 26.0%$.

- **How to Read t-Statistics and p-Values:** In regression tables, the number in parentheses next to an estimate is the **t-statistic** (how many standard errors the estimate sits away from zero). As a rule of thumb, $|t| \geq 2.0$ indicates a statistically reliable pattern ($p < 0.05$) that is unlikely to be a fluke of particular outlets or weeks. When we test multiple outcomes at once, we report **Holm-adjusted p-values** (`Holm p`), a strict penalty that guards against false alarms ("cherry-picking").
- **Where the Heavy Math Lives:** All core substantive tables and theoretical tests appear in **Sections 1 through 6**. A **Plain-Language Glossary** appears immediately after Section 6, and the formal econometric equations, two-way clustering formulas, spline knots, and 10-gate diagnostic details are preserved in **Appendices A, B, and C** so that general readers can read the main narrative smoothly while quantitative methodologists can inspect every technical specification.

Abstract

Digital communication scholarship widely concludes that negative emotional tone and moralized conflict amplify online engagement. Yet treating clicks, shares, upvotes, and raw comment totals as a single behavioral quantity obscures a foundational question in democratic theory: does affective news framing foster broad, evidence-grounded citizen dialogue, or does it merely trigger shallow reactions and polarized exchanges? Using a pre-registered two-stage hurdle design across M = 85 non-financial digital journalism outlets from January 2021 through December 2025, we separate initial crowd selection (N = 3653 outcome-blind submissions drawn from N = 220696 journalism posts) from 48-hour recursive conversation geometry (N = 2114 nested comment trees containing $226317 valid comments). Four open-weight language model families blindly scored 5398 unique headlines (21592 live evaluations) on Headline Negativity and Moral-Conflict and Blame Framing, achieving two-way absolute-agreement intraclass correlations of 0.9665 and 0.9650 in Stage 1 and 0.9638 and 0.9623 in Stage 2. Absorbing outlet, quarter, submission slot, and beat fixed effects alongside cubic splines of comment volume, we establish four findings. First, although negativity and moral conflict ignite initial first-comment reactions (+3.06 and +3.00 percentage points, $p = 0.0129$ and $p = 0.0211$) and crowd upvotes, neither predicts crossing the 15-comment front-page discussion hurdle ($t = +1.52$ and $t = -0.51$, Holm $p \geq 0.3946$). Second, within active discussions, non-conflictual negative valence is dead zero across all five deliberative metrics (Holm $p = 1.0000$), whereas moral-conflict framing shifts all five past multiplicity control ($p \leq 0.0177$). Third, exact additive tree decomposition proves that moral-conflict headlines induce *agonistic dyadic closure*: 80.44% (+1.09 percentage points, $t = +3.13$, $p = 0.0024$) of conflict's +1.35 percentage-point reciprocity gain ($t = +2.82$, Holm $p = 0.0240$) reflects repeat two-person ping-pong within locked-in pairs, while unique first-time dyad breadth stalls (+0.26 percentage points, $t = +1.05$), speakers per comment contract (-1.74 percentage points, $t = -3.34$), Shannon entropy drops (-0.0029, $t = -2.42$), and external documentary links fall (-0.73 percentage points, $t = -2.95$, Holm $p = 0.0125$). Fourth, readers meet news paywalls with a 5.38-fold surge in archive mirrors (1.089% versus 0.202%, regression-adjusted slope +0.0092,

t = +5.73), neutralizing paywall moderation (Holm p = 1.0000). Chronological 2024–2025 forecasts confirm that multi-judge semantic scores outperform structural controls and bag-of-words dictionaries ($p \leq 0.0342$).

1. Introduction

Few empirical regularities in contemporary communication science appear more settled than the dominance of negativity and moral conflict in digital attention markets. Psychophysiological and experimental studies show that citizens monitor threatening or adverse news more attentively than positive reporting (Soroka, 2014; Soroka & McAdams, 2015; Soroka, Fournier, & Nir, 2019; Trussler & Soroka, 2014), while large-scale field experiments demonstrate that negative headline words systematically elevate click-through rates (Robertson et al., 2023). Simultaneously, social-diffusion studies rooted in framing theory (Entman, 1993; Scheufele, 1999; Semetko & Valkenburg, 2000) show that moral-emotional expressions and out-group animosity substantially boost sharing and reaction totals on digital platforms (Brady et al., 2017; Rathje, Van Bavel, & van der Linden, 2021; Iyengar et al., 2018). Together, these literatures have fostered a *single-metric engagement paradigm* in which negative valence and conflictual framing are treated as unified accelerators of online civic participation.

From the perspective of public-sphere theory, however, collapsing clicks, upvotes, shares, and raw comment volume into an undifferentiated "engagement" index conflates low-cost attentional capture with democratic deliberation. Journalism scholars have long documented persistent "choice gaps" and "social news gaps" separating what editors prioritize, what readers click, and what audiences amplify socially (Boczkowski & Peer, 2011; Bright, 2016; Van Aelst et al., 2017). Recent feed-ranking experiments similarly reveal a wedge between behavioral interaction volume and democratic quality: divisive posts inflate clicks and comments while lowering user satisfaction and heightening partisan hostility (Milli et al., 2025; Piccardi et al., 2025). Normative accounts of the public sphere insist that the democratic value of citizen talk depends not on crowd noise, but on conversational structure and epistemic grounding—whether participants engage in reciprocal multi-turn dialogue, maintain broad speaker inclusion rather than two-person shouting matches, and substantiate contested claims with verifiable external sources (Habermas, 2006; Graham & Witschge, 2003; Stromer-Galley, 2007; Frieß & Eilders, 2015).

Once shallow crowd amplification is distinguished from downstream conversation architecture, core communication theories generate competing predictions. If negative and conflictual headlines primarily elicit expressive signaling or aversive withdrawal (Noëlle-Neumann, 1974; Mutz, 2002), they should inflate shallow reactions while crowding out reciprocal, evidence-based dialogue. Conversely, theories of affective intelligence (Marcus, MacKuen, & Neuman, 2011) and agonistic pluralism (Papacharissi, 2004; Karppinen, Moe, & Svensson, 2008; Muddiman & Stroud, 2017; Rossini, 2020) contend that conflict disrupts passive consensus and mobilizes deeper counter-argumentation. A third perspective—forum-design theory (Wright & Street, 2007; Halpern & Gibbs, 2012; Jaidka et al., 2021)—predicts that moderated, non-personalized forum architecture decouples downstream conversation topology from headline framing altogether ($\beta_2 = 0$).

Testing these rival mechanisms requires overcoming four pervasive empirical pitfalls: (1) viral-survival collider bias when sampling only active discussion threads (Heckman, 1979); (2) construct conflation between non-conflictual tragedy or threat (Soroka, Fournier, & Nir, 2019) and adversarial blame attribution (Entman, 1993); (3) mechanical scaling of tree depth and turn counts with comment volume, compounded by author self-threads and paywall-mirror hyperlinks; and (4) institutional access barriers from digital news paywalls (Myllylahti, 2013; Sjøvaag, 2015; Benson, 2006).

We address all four challenges via a pre-registered two-stage investigation strictly within communication and media science—excluding all financial press, corporate earnings stories, and market variables—across M = 85 journalism outlets over 2021–2025. Pairing an outcome-blind Stage-1 submission panel (N = 3653 from N = 220696 clean journalism submissions) with full 48-hour recursive discussion trees in Stage 2 (N = 2114 threads containing $226317 comments), and scoring headlines blindly across four independent open-weight language model families (21592 live ballots), we uncover a sharp structural dissociation. Headline negativity and conflict spark initial zero-to-one reactions and upvotes but stall at the 15-comment community gatekeeping hurdle. Inside active discussions, pure negative valence is completely inert across all five deliberative metrics, whereas moral-conflict framing drives *agonistic dyadic closure*—elevating multi-turn reciprocity almost entirely (80.44%) through repeat two-person ping-pong among fewer speakers while eroding external documentary evidence citing. Meanwhile, a 5.38$-fold surge in communal archive mirrors neutralizes paywall information starvation, demonstrating grassroots *networked epistemic repair* (Barzilai-Nahon, 2008).

---

2. Theoretical Framework and Competing Hypotheses

2.1 Psychophysiological Negativity Versus Moral-Conflict Framing: Why Tragedy Is Not Blame

Two distinct mechanisms explain why adverse and contentious news captures audience attention, yet empirical studies routinely merge them into a single sentiment dimension. The first is *psychophysiological negativity bias*. Cross-national physiological experiments show that human autonomic arousal responds more strongly to negative news—disease outbreaks, accidents, environmental or civic deterioration—than to positive news, regardless of self-reported preferences (Soroka, 2014; Soroka & McAdams, 2015; Soroka, Fournier, & Nir, 2019; Trussler & Soroka, 2014). Analyzing randomized headline experiments, Robertson et al. (2023) confirmed that negative words significantly raise click-through rates. Crucially, threat surveillance requires no human culprit: a natural disaster or medical hazard carries high negative valence without interpersonal blame.

The second mechanism is *moral-conflict framing and blame attribution*. Entman (1993) conceptualized framing around problem definition, causal attribution, moral evaluation, and treatment recommendation (Scheufele, 1999), and newsroom analyses identify conflict and responsibility attribution as pervasive journalistic templates (Semetko & Valkenburg, 2000). In networked environments, moralized conflict engages group identity and moral indignation: moral-emotional words and out-group animosity capture visual attention and drive social sharing (Brady et al., 2017; Rathje, Van Bavel, & van der Linden, 2021; Iyengar et al., 2018).

Yet because clicking, sharing, and deliberating obey distinct logics (Boczkowski & Peer, 2011; Bright, 2016; Neuman et al., 2014), whether valence negativity and moral conflict enrich or degrade downstream conversation remains an open theoretical clash.

> ***Intuitive Example from Our Corpus — Pure Threat/Tragedy vs. Moral-Conflict Blame:*** *Consider two real stories in our Stage-2 deliberative tree panel: -* ***High Negativity, Low Moral Conflict (Pure Threat/ Tragedy, C_1):*** *"A suicide crisis among veterinarians" (BBC, story* `37844225` *). All four language-model judges rated its Headline Negativity very high ($0.887 out of 1.00) because the subject is a painful public-health tragedy, but rated its Moral-Conflict and Blame Framing near zero (0.075) because the headline does not accuse an adversary of moral wrongdoing. Across 401 valid 48-hour comments, readers shared sympathy, personal experiences, and occupational analyses: its repeat two-person ping-pong rate remained low (0.0880, or 8.80%$). -* ***High Moral Conflict and Blame (C_2):*** *"Anti-cybercrime laws are being weaponized to repress journalism" (Columbia Journalism Review, story* `45792209` *, Negativity $0.713, Conflict 0.713) or "Supreme Court will decide whether web designer has right to turn away gay couple" (Los Angeles Times, story* `30437122` *, Negativity 0.325, Conflict 0.775). Here, the headline explicitly frames an institutional or constitutional clash between opposing sides. In these threads (95 and 186 comments inside 48 hours), total A \to B \to A reciprocity jumped above 41%, driven heavily by repeat two-person ping-pong (32.05% and 23.03%) while external documentary link rates dropped to 2.11% and 6.45%$.*

2.2 Public-Sphere Crowding-Out, Agonistic Mobilization, and Forum Affordances

Three theoretical traditions yield rival predictions for Stage-2 conversation geometry.

Habermasian Crowding-Out (H_1). Normative deliberative theory grounds democratic legitimacy in reciprocal listening, inclusive speaker participation, and reasoned documentary justification (Habermas, 2006). Empirical deliberation schemes operationalize these norms via multi-turn interpersonal reciprocity (A \to B \to A), speaker diversity, and external evidence citing (Graham & Witschge, 2003; Stromer-Galley, 2007; Frieß & Eilders, 2015). Combined with spiral-of-silence and cross-cutting exposure research showing that adversarial climates chill broader participation (Noëlle-Neumann, 1974; Mutz, 2002), affective and conflictual headlines should boost Stage-1 volume ($\beta_1 > 0$) while crowding out Stage-2 reciprocity, nesting, speaker breadth, and documentary citing into shallow depth-1 venting ($\beta_2 < 0$; Milli et al., 2025).

Agonistic Deliberative Mobilization (H_2). Critics of consensus deliberation argue that impassioned contestation and sharp disagreement are catalysts rather than enemies of democratic talk (Papacharissi, 2004; Karppinen, Moe, & Svensson, 2008). Empirical work separating substantive partisan or issue conflict from personal intolerance shows that conflictual cues stimulate active rebuttals (Muddiman & Stroud, 2017; Rossini, 2020), while affective intelligence and polarized-crowd theories hold that threat and conflict motivate scrutiny

and counter-evidence search (Marcus, MacKuen, & Neuman, 2011; Shi et al., 2019). Under H_2, framing elevates both Stage-1 entry ($\beta_1 > 0$) and Stage-2 reciprocal argumentation and evidence citing ($\beta_2 > 0$).

Valence-versus-Conflict Bifurcation (H_3) and Platform Decoupling (H_0). Because uncontested tragedy elicits consensual sympathy whereas blame attribution posits competing sides (Entman, 1993; Semetko & Valkenburg, 2000), H_3 predicts systematic divergence between Moral-Conflict Framing and pure Headline Negativity ($\beta_{2,conf} \neq \beta_{2,neg}$). Conversely, institutional design theory contends that threaded architecture, community moderation, and journalistic beat norms govern how citizens converse once inside a forum, rendering Stage-2 geometry invariant to headline framing ($\beta_2 \approx 0$; Wright & Street, 2007; Halpern & Gibbs, 2012; Jaidka et al., 2021).

2.3 Journalistic Fields, Paywalls, and Strong-Inference Matrix

Journalistic framing operates across differentiated fields—from national public-affairs wires to specialist science, technology, and culture beats (Benson, 2006)—and across open-access versus subscription business models (Myllylahti, 2013; Sjøvaag, 2015). Under our paywall information-starvation hypothesis (H_4), restricting article text behind paywalls forces non-subscribers to react to headline cues alone, steepening framing-induced deliberative erosion ($\gamma_{pw} < 0$) unless offset by networked counter-gatekeeping (Barzilai-Nahon, 2008; Carlson, 2015).

We therefore test four research questions—Stage-1 crowd breakthrough (RQ_1), Stage-2 deliberative geometry (RQ_2), valence-versus-conflict bifurcation (RQ_3), and paywall moderation (RQ_4)—summarized in Table 1.

Hypothesis	**Canonical Lineage**	**Stage-1 Breakthrough (β_1)**	**Stage-2 Size-Adjusted Deliberation & Evidence (β_2)**	**Predicted Equilibrium**
H_1: Crowding-Out	Robertson et al. (2023); Brady et al. (2017); Habermas (2006); Milli et al. (2025)	$\beta_1 > 0$	$\beta_2 < 0$ (reciprocity, nesting, & evidence fall)	*Clicks Without Conversation*: Shallow volume crowds out reciprocal, evidence-based talk.
H_2: Agonistic Mobilization	Marcus et al. (2011); Papacharissi (2004); Muddiman & Stroud (2017); Rossini (2020)	$\beta_1 > 0$	$\beta_2 > 0$ (reciprocity, depth, & evidence rise)	*Productive Contestation*: Conflict mobilizes multi-turn rebuttals and counter-evidence.
H_3: Valence-Conflict Split	Soroka et al. (2019) vs. Entman (1993);	$\beta_{1,conf} \neq \beta_{1,neg}$	$\beta_{2,conf} \neq \beta_{2,neg}$	*Valence Is Not Conflict*: Blame/ contestability diverges

Hypothesis	Canonical Lineage	Stage-1 Breakthrough (β_1)	Stage-2 Size-Adjusted Deliberation & Evidence (β_2)	Predicted Equilibrium
	Semetko & Valkenburg (2000)			sharply from pure tragedy.
H_0: Affordance Decoupling	Wright & Street (2007); Halpern & Gibbs (2012); Frieß & Eilders (2015); Jaidka et al. (2021)	$\beta_1 \geq 0$	$\beta_2 \approx 0$ (Holm $p \geq 0.05$)	*Architecture Over Headlines*: Forum design insulates talk from headline framing.
H_4: Paywall Starvation	Myllylahti (2013); Sjøvaag (2015); Benson (2006); Barzilai-Nahon (2008)	Entry friction	$\gamma_{pw} < 0$ on paywalled domains	*Headline-Only Reacting*: Paywalls worsen distortion unless repaired by communal mirrors.

3. Pre-Registered Two-Stage Hurdle Data and Research Design

3.1 Curated Journalism Universe and Pre-Data Selection Protocol

Before locking our confirmatory design in a standalone cryptographic commit (`131fa8a`) prior to full-panel ingestion, we audited five pure-communication designs using live archival probes. Four alternatives were rejected pre-data due to construct censoring or selection bias: within-thread incivility testing suffered $80.00% floor-censoring ($\leq 0.05$) under anti-insult moderation; primary-versus-legacy news dyads yielded only a 17.24% match rate under platform duplicate suppression; encyclopedia talk-page gatekeeping exhibited 54.17% acute zero-censoring; and preprint abstract hedging-versus-hype faced viral-survival collider bias because over 90%$ of raw preprints draw zero to two comments.

Locking the digital journalism design on Hacker News enabled a clean *two-stage hurdle architecture* across complete, timestamped conversation trees without personalized feed confounding. Enforcing $100% communication-science purity, we excluded all financial and business outlets and applied a deterministic headline filter removing corporate earnings, share-price, initial public offering, securities-filing, and cryptocurrency market stories. Our curated registry comprises M = 85 qualifying non-financial journalism outlets (58 open-access, 27$ paywalled) across five beats: (1) national, international, and public-service wires/ broadcasters (*Reuters*, *Associated Press*, *BBC*, *NPR*, *The Guardian*, *The New York Times*, *The Washington Post*, *Politico*); (2) investigative, legal, and policy reporting (*ProPublica*, *The Markup*, *SCOTUSblog*, *Lawfare*, *Bellingcat*); (3) science, environment, and health journalism (*Nature*, *Science*, *Quanta Magazine*, *STAT*,

Scientific American); (4) technology-in-society journalism (*Ars Technica*, *Wired*, *The Verge*, *MIT Technology Review*, *404 Media*); and (5) culture and ideas magazines (*The Atlantic*, *The New Yorker*, *Harper's*, *Aeon*, *Vox*, *Slate*).

Adaptive time-sliced queries across January 1, 2021 through December 31, 2025 (< 900 hits per slice, > 10 months matured so right-censoring equals \$0%) scanned N = 221497 raw submissions. Dropping 294 host mismatches, 14 platform meta-posts, and 493 market headlines yielded N = 220696 clean journalism submissions across all M = 85 outlets (N = 23453 reaching ≥ 15\$ comments), from which we drew two panels:

1. **Stage-1 Unconditional Panel (N = 3653, G = 85 outlets, G = 262 weeks):** A seeded stratified random draw across outlet-by-year cells over all submissions (≥ 0 comments), with inverse-probability weights summing exactly to \$220696.00 (398 stories reaching ≥ 15\$ comments automatically entered Stage 2).
2. **Stage-2 Deliberative Tree Panel (N = 2114, G = 85 outlets, G = 261 weeks):** Combining those \$398 stories with stratified discussed draws yielded 2251 candidate trees (100% retrieval; weights summing to 23453). Retaining threads with ≥ 10 valid comments inside 48 hours produced N = 2114 recursive trees containing 226317 valid 48-hour comments (235510\$ over 14 days).

> ***Plain-Language Method Box 1: Why Use a Two-Stage "Hurdle" Design—and Why Are Primary Regressions Unweighted While Inverse-Probability Weights (IPW) Are Tested Separately? - Why Two Stages?*** *Imagine trying to study whether dramatic movie titles make audiences stay in their seats for post-screening discussions, but only collecting data inside theaters where at least 15 people stayed to talk. If dramatic titles affect which screenings attract 15 people in the first place, studying only the surviving discussions creates what statisticians call* ***collider or selection bias*** *(Heckman, 1979). Stage 1 observes all submissions—including the \$56.75% of stories that receive zero comments and the 89.10%\$ that never reach 15 comments—while Stage 2 analyzes the internal conversation geometry of threads that actually cross the discussion threshold. -* ***What Are Inverse-Probability Weights (IPW), and Why Are the Primary Models Unweighted?*** *In our full five-year universe of \$220696 journalism posts, a handful of high-volume outlets (such as The New York Times, The Guardian, or Ars Technica) account for thousands of submissions, while smaller investigative or science magazines (such as The Markup or Quanta Magazine) account for far fewer. If we drew a simple uniform random sample, small outlets would barely appear. Instead, we used* ***stratified sampling*** *across every* `(outlet, year)` *cell so that all 85 journalism domains are well-represented in our 3653-story and 2114-tree panels. For every sampled story, we recorded its exact* ***Inverse-Probability Weight (IPW)****—how many stories in the full 220696-story universe that sampled story represents (\sum IPW = 220696.00 in Stage 1; \sum IPW = 23453.00\$ in Stage 2). -* ***Following standard econometric practice when outlet and time fixed effects are already included in the regression****, our primary models estimate* ***unweighted within-outlet slopes*** *(so that giant newspapers do not swamp the statistical precision of the 85-outlet panel), and we then explicitly test* ***IPW universe-weighted regressions and Heckman selection-corrected models*** *in our falsification gates (Gates 3 and 7; see Section 4.3 and Appendix B). As the Singapore sister-lab external*

> *audit noted, both unweighted and IPW-weighted specifications yield the exact same substantive conclusions (in fact, IPW weighting slightly strengthens Moral Conflict's effect on Stage-2 reciprocity from +0.0135, t = +2.82 to +0.0179, t = +3.47 and +0.0202, t = +3.33). - (Audit Provenance Note on Raw Hits): Across all raw cache files, shipped attrition diagnostics, verification logs, and site prompts, the exact raw Algolia hit count scanned is N = 221497 ($221497 - 294 - 14 - 493 = 220696 clean stories); an early draft of our standalone self-audit memo momentarily mis-transcribed the raw hit count as 238311 before being reconciled to 221497$.*

3.2 Blind Four-Family Semantic Annotation and Shrout–Fleiss Reliability

All $5398 unique headlines were scored blindly at temperature 0.0 (seeing *only* headline text, stripped of outlet, date, points, and comments) by four independent open-weight model families—Llama 3.3 70B, Qwen 2.5 72B, DeepSeek V3, and Gemma 3 27B—producing 21592 live JSON evaluations (0 fallbacks) on [0.00, 1.00] for **Headline Negativity** (N_i: harm, crisis, loss, threat, or adverse outcomes without requiring blame) and **Moral-Conflict and Blame Framing** (C_i$: interpersonal, institutional, or political conflict, accusation, or moral wrongdoing).

> ***Plain-Language Method Box 2: Why Use Four Independent AI Judges Instead of Word Lists—and What Is an Intraclass Correlation (ICC)? - Why Word Lists Fail on Headlines:*** *Traditional automated text analysis counts words from pre-made lists (for example, counting words like crisis, crash, attack, or slam). While we include two comprehensive word-list dictionaries ($2487 negative words and 199$ conflict words) as control variables in every regression, word lists cannot tell whether "A suicide crisis among veterinarians" is a tragic health story (Negativity) or a partisan political attack (Moral Conflict), nor can they recognize moral conflict when a headline uses calm legal language like "Supreme Court will decide whether web designer has right to turn away gay couple." - **Why Four Different AI Model Families?** Relying on a single AI model (such as GPT or Claude alone) risks baking that specific model's quirks into the measurement. By querying **four distinct open-weight model families built by four different organizations** (Meta's Llama 3.3 70B, Alibaba's Qwen 2.5 72B, DeepSeek V3, and Google's Gemma 3 27B) at `temperature = 0.0` (maximum determinism) and averaging their four independent ballots (**Joint Evaluation, or JEV**), we treat the four models like a panel of four independent human coders. - **What Does ICC(2,4) > 0.96 Mean?** The **Intraclass Correlation Coefficient (ICC)** (Shrout & Fleiss, 1979) measures how closely multiple judges agree on the exact numerical score of each headline, from $0.00 (random noise) to 1.00 (perfect identity). In social science content analysis, an ICC above 0.75 is considered good and above 0.90 is exceptional. Across our 5398 headlines, the four-judge panel achieves **two-way absolute agreement ICC(2,4) between 0.9623 and 0.9665** (and single-judge reliability ICC(2,1) ≥ 0.8645$). Even more importantly, in **Gate 1** and **Check E** we re-run our regressions using*

> *each individual AI judge alone and dropping one judge at a time to prove that no single AI model drives our findings.*

Table 2 reports textbook two-way Shrout and Fleiss (1979) reliability. Four-judge absolute agreement ICC(2,4) reaches $0.9665 and 0.9650 in Stage 1 and 0.9638 and 0.9623 in Stage 2 (single-rater ICC(2,1) ≥ 0.8645, pairwise r ∈ [0.8684, 0.8886]). Negativity and Conflict correlate at r = +0.7533 in Stage 1 and r = +0.7176 in Stage 2, yielding modest variance inflation factors (2.3527 and 2.4105 in Stage 1; 2.1736 and 2.2687 in Stage 2). Pre-registered 0.50-midpoint typology counts in Stage 2 are 1198 Neutral/Constructive (C_0), 315 Pure Threat/Tragedy (C_1), 412 Moral-Conflict/Blame (C_2), and 189 Non-Negative Debate (C_3; Stage-1 counts: 2342, 441, 568, and 302$).

Panel & Construct	N	ICC(1,4)	ICC(2,4)	ICC(3,4)	ICC(2,1)	Pairwise r	Mean (SD)	Exact Zero
Stage 1: Headline Negativity	$3653	0.9664	0.9665	0.9682	0.8782	0.8886	0.2944 (0.2734)	20.67% $
Stage 1: Moral-Conflict & Blame	$3653	0.9649	0.9650	0.9655	0.8732	0.8768	0.2465 (0.2741)	31.37% $
Stage 2: Headline Negativity	$2114	0.9637	0.9638	0.9659	0.8694	0.8806	0.3460 (0.2750)	15.14% $
Stage 2: Moral-Conflict & Blame	$2114	0.9623	0.9623	0.9629	0.8645	0.8684	0.2903 (0.2817)	23.37% $

3.3 Conversation-Tree Outcomes and Econometric Specifications

Stage-1 outcomes are Discussed Breakthrough (≥ 15 comments, base rate $10.90%), Log Upvote Points (mean +1.8228, SD 1.1932), and Log Comment Volume (mean +0.8259, SD 1.3806). In Stage 2, we retain valid comments posted within 48 hours (172800 seconds; 6016 later comments pruned) and exclude 562 author self-replies (0.25%$) from interpersonal metrics across five outcomes: 1. **Dyadic Reciprocity Rate (48h, Co-Primary):** Share of interpersonal replies (d ≥ 2, distinct parent–child authors) where child author equals grandparent author (d ≥ 3) across two distinct citizens (A \to B \to A; mean +0.2117, SD 0.1053). 2. **Interpersonal Nested Reply Share (48h):** Share of valid 48-hour comments that are interpersonal nested replies rather than depth-1 root statements or self-appends (mean +0.7982, SD 0.1111). 3. **90th-Percentile Branch Depth (48h):** Upper-decile tree depth (mean +5.2674, SD 1.6128). 4. **External Evidence Link Rate (48h):** Share of HTML-unescaped 48-hour comments citing external URLs after stripping archive/

paywall mirrors, platform self-links, and article self-domain links (mean +0.0991, SD 0.0650), logging **Paywall Mirror Link Rate** separately. 5. **Speaker Shannon Entropy (48h):** Normalized turn-share entropy [0, 1] across unique commenters (mean +0.9548, SD 0.0269).

> ***Visualizing a Recursive Conversation Tree: What Do Our Five Stage-2 Metrics Actually Measure?*** *Imagine a news article is posted at the top of a forum (Depth 0: Story Root). Readers can either leave a standalone comment directly on the article (Depth 1) or reply directly underneath another citizen's comment (Depth 2, 3, 4...):*
>
> - ***[Story Root: News Article Headline]*** *(Depth 0)*
> - ***Comment 1 by Citizen A*** *(Depth 1: Standalone root reaction to the headline)*
>   - ***Comment 2 by Citizen B replying to Citizen A*** *(Depth 2: Interpersonal nested reply)*
>   - ***Comment 3 by Citizen A replying to Citizen B*** *(Depth 3: Completes a three-turn A \to B \to A reciprocal triad!)*
>   - ***Comment 4 by Citizen B replying to Citizen A*** *(Depth 4: Completes B \to A \to B, a* ***repeat ping-pong volley*** *within the same {A, B} pair!)*
> - ***Comment 5 by Citizen C*** *(Depth 1: Standalone root reaction citing an external scientific or legal source URL)*
> - ***Comment 6 by Citizen D*** *(Depth 1: Posts an archive mirror link so non-subscribers can read the full text)*
> - ***Metric 1 (Dyadic A \to B \to A Reciprocity Rate):*** *Among all person-to-person replies (Depth ≥ 2), what fraction complete a three-turn loop where the grandparent commenter replies back to the person who responded to them (Comment 3 is an A \to B \to A turn; Comment 4 is a second B \to A \to B turn within that same pair)? Across our $2114 threads, an average of* ***21.17%*** *of interpersonal replies are A \to B \to A$ reciprocal turns.*
> - ***Metric 2 (Interpersonal Nested Reply Share):*** *What fraction of all comments in the thread are person-to-person replies (Depth ≥ 2, excluding an author replying to their own comment) rather than people just shouting standalone Depth-1 reactions at the headline? On average,* ***$79.82%$*** *of comments are nested replies.*
> - ***Metric 3 (90th-Percentile Branch Depth):*** *How deep do the longest chains of back-and-forth exchange reach? Rather than looking only at the single deepest comment (which can be an outlier), the 90th-percentile depth measures the depth level reached by the deepest $10% of comments in the thread (averaging* ***5.2674$ levels deep****).*
> - ***Metric 4 (External Evidence Link Rate):*** *What percentage of comments bring outside documentary receipts—linking to scientific papers, court filings, government statistics, or other news reports*

> *(Comment 5)? Crucially, we **strip out paywall-bypass archive links (Comment 6), links back to Hacker News itself, and links to the same news domain** so that sharing an archive mirror is never confused with citing independent evidence. On average, **$9.91%$** of comments cite external documentary links.*
>
> - ***Metric 5 (Speaker Shannon Entropy):** Borrowed from information theory, this measures **equality of voice** on a $0.00 to 1.00 scale. If 100 different citizens each speak once, entropy is 1.0000$ (maximum diversity). If two loud commenters monopolize half the thread while everyone else speaks once, entropy drops.*

All models absorb outlet domain ($85), year-quarter, UTC day-by-six-hour slot, and URL beat fixed effects plus structural headline features and deterministic lexical negativity (2487 words) and conflict (199 words) dictionary rates. Stage-2 models include four-knot restricted cubic splines of \log(V_{48}) (knots 2.772589, 3.663562, 4.543295, 5.834811). Inference uses two-way `(domain, week)` clustering (Cameron, Gelbach, & Miller, 2011) and Holm (1979) multiplicity adjustments (K = 4 co-primary slopes and K = 5$ Stage-2 outcomes).

> ***Plain-Language Method Box 3: What Do Fixed Effects, Volume Splines, Two-Way Clustering, and Holm Correction Do? - Four-Way Fixed Effects (Comparing Apples to Apples):** You might worry that The New York Times naturally gets different comments than Nature, or that election quarters in 2024 look different from early 2021, or that stories posted on a Tuesday morning in the U.S. behave differently from stories posted on a Sunday night. **Fixed effects** mathematically subtract out the baseline average of every single news outlet ($85 domains), every calendar quarter (20 quarters), every day-of-week-by-6-hour posting window (28$ slots), and every URL section beat. That means our estimates compare **more-conflictual versus less-conflictual headlines published by the exact same news outlet, in the same topic beat, during the same calendar quarter and time of day**. - **Restricted Cubic Splines of Comment Volume (\log V_{48}):** A thread with $300 comments naturally has deeper branches than a thread with 16 comments simply because there are more comments. Instead of assuming a rigid straight-line relationship between thread size and tree structure, **four-knot restricted cubic splines** fit a smooth, flexible curve across thread sizes (splitting at roughly 16, 39, 94, and 342$ comments). This ensures we are comparing the shape of conversations among threads of the **exact same size**. - **Two-Way Clustered Standard Errors (Cameron, Gelbach, & Miller, 2011):** Standard textbook formulas assume every news story is completely independent. In reality, ten stories from Reuters share reader habits, and thirty stories posted during the same week share the week's news mood. Clustering simultaneously by **outlet (G = 85)** and **calendar week (G = 262 or 261)** widens our error bars to account for both outlet correlation and weekly news shocks. - **Holm Multiplicity Correction (Holm, 1979):** If you flip a coin enough times—or test enough regression coefficients—you will eventually get a "statistically significant" $p < 0.05$ by sheer luck. Because we pre-registered K = 4 co-primary tests across Stages 1 and 2 and K = 5 deliberative tree*

> *outcomes in Stage 2, the **Holm step-down procedure** multiplies the smallest p-value by $5 (or 4), the next smallest by 4 (or 3$), and so on. A result passes only if it survives this severe penalty.*

4. Empirical Results

4.1 Stage-1 Crowd Gatekeeping: Upvotes Without Front-Page Breakthrough

Table 3 reports Stage-1 estimates (N = 3653). On Discussed Breakthrough (≥ 15 comments), neither Headline Negativity (+0.0158, SE = 0.0104, t = +1.52, p = 0.1315, K = 4 Holm p = 0.3946) nor Moral-Conflict Framing (-0.0059, SE = 0.0115, t = -0.51, p = 0.6079, Holm p = 0.6199) is statistically significant. Standalone models (t = +1.65 and +0.70), inverse-probability universe-weighted models (t = +0.21 and +0.66), and logit estimates (z = +1.52 and +0.10) confirm this front-page breakthrough null (omitting lexical controls yields only a faint surface-lexical negativity slope of +0.0203, t = +2.01, p = 0.0473).

Stage-1 Outcome (N = 3653)	Joint Neg (z)	Joint Conflict (z)	Standalone Neg	Standalone Conf	Pure Threat (C_1)	Moral Blame (C_2)	Debate (C_3)
Breakthrough (≥ 15)	+0.0158 (+1.52)	-0.0059 (-0.51)	+0.0120 (+1.65)	+0.0057 (+0.70)	+0.0231 (+1.24)	+0.0016 (+0.09)	+0.0257 (+1.03)
Log Upvote Points	+0.0579 (+1.66)	+0.0872^* (+2.29)	+0.1129** (+4.67)	+0.1299** (+4.88)	+0.1277^* (+2.29)	+0.2387** (+3.89)	+0.2675** (+2.69)
Log Comment Volume	+0.0693 (+1.66)	+0.0459 (+1.14)	+0.0983** (+3.38)	+0.0971** (+3.46)	+0.1535^* (+2.06)	+0.1237^* (+2.00)	+0.2516^* (+2.25)

Notes: Four-way fixed effects plus structural/lexical controls; two-way clustered t-statistics in parentheses (G = 85, G = 262). ^*p < 0.05*, *p < 0.01 (unadjusted; K = 4 Holm p = 0.3946 and 0.6199 on Breakthrough).

By contrast, continuous crowd volume responds strongly at lower margins: standalone Negativity and Conflict elevate Log Upvote Points (+0.1129, t = +4.67, p = 0.000011; +0.1299, t = +4.88, p = 0.000005) and Log Comments (+0.0983, t = +3.38, p = 0.0011; +0.0971, t = +3.46, p = 0.0008). Jointly, Moral-Conflict Framing significantly lifts Log Upvote Points (+0.0872, SE = 0.0381, t = +2.29, p = 0.0246) and Poisson comment counts (+0.2869, z = +2.16, p = 0.0311), with Moral Blame (+0.2387, p = 0.0002) and Non-Negative Debate (+0.2675, p = 0.0086) leading neutral headlines.

> ***What Do the Stage-1 Effect Sizes Mean in Practice?** Why do we see a positive effect on Log Upvote Points and first comments, yet a flat zero on reaching 15 comments? Look at the numbers in concrete terms: - Across all $3653 submissions, **56.75% receive zero comments**, 43.25% receive at least 1*

> *comment, and only **10.90% (398$ stories) cross the 15-comment breakthrough threshold**. - When a headline ramps up Negativity or Moral Conflict by one standard deviation, it nudges the probability of getting **at least one comment** up by +3.06 and +3.00 percentage points (moving from a $43.25% chance of a first comment to roughly 46.3%), and Moral Conflict lifts upvote points by roughly +8.72% (\Delta \log(points) = +0.0872 per SD, or +23.87% for Moral Blame headlines in cell C_2 compared to neutral headlines in C_0$). - Yet when we ask whether that headline pushes the story across the **15-comment community discussion threshold**, the effect disappears (+1.58 percentage points for Negativity, t = +1.52; -0.59 percentage points for Conflict, t = -0.51). In other words: emotional and conflictual headlines are good at provoking **a couple of initial drive-by comments and passive upvotes**, but in a forum without an algorithmic outrage feed, community gatekeeping prevents those drive-by sparks from turning into sustained front-page conversations.*

4.2 Stage-2 Deliberative Architecture: Valence Null Versus Five-Dimension Conflict Shift

Table 4 presents Model 2 and Model 3 results across N = 2114 recursive 48-hour conversation trees ($226317 comments), controlling for four-knot cubic splines of \log(V_{48})$.

First, **Headline Negativity conditional on Moral Conflict is precisely null across all five deliberative outcomes (K = 5 Holm p = 1.0000 on five of five)**: Dyadic Reciprocity (+0.0035, SE = 0.0034, t = +1.02, p = 0.3100, K = 4 Holm p = 0.6199), Interpersonal Nested Reply Share (-0.0018, t = -0.60), 90th-Percentile Branch Depth (+0.0117, t = +0.27), External Evidence Link Rate (+0.0017, t = +0.74), and Speaker Shannon Entropy (-0.0004, t = -0.39), plus Mean Depth (-0.0012, t = -0.06) and Log Comment Words (-0.0110, t = -0.95). Standalone negativity correlations with reciprocity (+0.0118, t = +4.51) vanish completely once moral blame is separated.

Stage-2 48h Outcome (N = 2114)	Joint Neg (z)	Joint Conflict (z)	K = 5 Holm p	Conflict β_{std}	Wald (Conf - Neg)	Moral Blame (C_2)	Debate (C_3)
Dyadic A \to B \to A Reciprocity	+0.0035 (+1.02)	+0.0135** (+2.82)	$0.0125	+0.1286 SD	+0.0101 (0.1829)	+0.0246^{**} (+2.73)	+0.0212^* (+2.20$)
Interpersonal Nested Share	-0.0018 (-0.60)	+0.0127** (+3.76)	$0.0013	+0.1145 SD	+0.0146^* (0.0158)	+0.0204^{**} (+3.16)	+0.0179^{**} (+2.70$)
90th-Pct Branch Depth	+0.0117 (+0.27)	+0.1812** (+3.80)	$0.0013	+0.1123 SD	+0.1695^* (0.0395)	+0.3487^{**} (+3.36)	+0.2694^{**} (+2.99$)
External Evidence Link Rate	+0.0017 (+0.74)	-0.0073** (-2.95)	$0.0125	-0.1121 SD	-0.0090^* (0.0500)	-0.0031 (-0.91)	-0.0054 (-1.18$)

Stage-2 48h Outcome (N = 2114)	Joint Neg (z)	Joint Conflict (z)	K = 5 Holm p	Conflict \beta$_{std}$	Wald (Conf - Neg)	Moral Blame (C_2)	Debate (C_3)
Speaker Shannon Entropy	-0.0004 (-0.39)	-0.0029^* (-2.42)	$0.0177	-0.1063 SD	-0.0025 (0.2048)	-0.0040 (-1.84)	-0.0028 (-0.97$)

Notes: Four-way fixed effects, four-knot \log(V_{48}) splines, and structural/lexical controls; two-way `(domain, week)` t-statistics (G = 85, G = 261) or Wald p-values in parentheses. Primary K = 4 Holm p for Reciprocity Conflict equals $0.0240; all five Negativity Holm p-values equal 1.0000$.

Second, **Moral-Conflict and Blame Framing shifts all five deliberative dimensions past K = 5 Holm multiplicity control (p ≤ 0.0177), splitting structural engagement against epistemic quality**. A one-SD increase in Moral Conflict *increases* Dyadic Reciprocity Rate by +0.0135 (+4.81 percentage points across [0, 1], SE = 0.0048, t = +2.82, p = 0.0060, K = 4 Holm p = 0.0240, K = 5 Holm p = 0.0125; total-comment-denominator reciprocity +0.0137, t = +3.28), *increases* Interpersonal Nested Reply Share by +0.0127 (SE = 0.0034, t = +3.76, Holm p = 0.0013), and *deepens* 90th-Percentile Branch Depth by +0.1812 levels (SE = 0.0476, t = +3.80, Holm p = 0.0013; Mean Depth +0.1010, t = +4.27, p = 0.000050). Yet Moral Conflict simultaneously *reduces* External Evidence Link Rate by -0.0073 per SD (-2.59 percentage points across [0, 1] off the +0.0991 sample mean; SE = 0.0025, t = -2.95, p = 0.0042, Holm p = 0.0125) and *reduces* Speaker Shannon Entropy by -0.0029 (SE = 0.0012, t = -2.42, p = 0.0177, Holm p = 0.0177), while comment length stays flat (+0.0166, t = +1.15, p = 0.2528). Continuous Wald tests (Conflict - Negativity) confirm significant divergence on Nested Share (p = 0.0158), 90th-Percentile Depth (p = 0.0395), Mean Depth (p = 0.0121), and Evidence Links (p = 0.0500).

> ***Translating Table 4 into Substantive Terms: Why Standardised Effects of ± 0.11 to 0.13 SD Are Large for Downstream Conversation Trees*** *At first glance, a reader unfamiliar with conversation-tree data might ask: "Is a +1.35 percentage-point shift in reciprocity or a -0.73 percentage-point shift in external links per standard deviation big or small?" Three benchmarks show why these shifts are substantively consequential: 1.* ***Comparing Against the Baseline Rates:*** *Across all $2114 threads, only* ***9.91% of comments (0.0991) contain an external documentary link****. A one-SD increase in Moral Conflict reduces external evidence linking by -0.73 percentage points (≈ 7.4% relative reduction per SD), and moving from a non-conflictual headline (C_i = 0) to a full moral-conflict headline (C_i = 1) reduces external evidence linking by* ***-2.59 percentage points—cutting out more than one-quarter (26.1%$) of the baseline rate of documentary fact-citing****! 2.* ***Standardized Effect Sizes (\beta$_{std}$):*** *Expressed in standard deviations of the outcome variables (the standardized conflict column in Table 4), a one-SD increase in headline moral conflict shifts all five downstream conversation dimensions by* ***+0.1286 SD*** *(reciprocity),* ***+0.1145 SD*** *(nested reply share),* ***+0.1123 SD*** *(90th-percentile depth),* ***-0.1121 SD*** *(evidence links), and* ***-0.1063 SD*** *(speaker entropy). Because a news headline is just a 10-to-15-word cue read before clicking*

> *an article and before hundreds of independent citizens interact over 48 hours, finding that headline conflict alone shifts the downstream 48-hour topology of a 100-comment tree by more than a tenth of a standard deviation across every single structural dimension—after stripping out outlet, beat, time, word-list dictionaries, and thread volume—demonstrates a powerful structural imprint. 3.* ***Why Naive Studies Mistook Negativity for an Accelerator:*** *Notice that if you put Headline Negativity into a regression by itself without controlling for Moral Conflict (Standalone Negativity), Negativity appears to strongly increase reciprocity (+0.0118, t = +4.51, p = 0.000020). The moment Moral Conflict is entered alongside it (Joint Negativity), Negativity's effect collapses to +0.0035 (t = +1.02, Holm p = 1.0000). Earlier studies that measured only "negative sentiment" were picking up the hidden shadow of moral conflict and blame.*

4.3 Paywall Moderation, Robustness Gates, and Out-of-Sample Forecasting

Testing H_4 (Model 4), paywalled outlets (M = 27) trigger a $5.38-fold jump in Paywall Mirror Link Rate over open-access outlets (M = 58): 1.089% versus 0.202% (regression slope +0.0092, SE = 0.0016, t = +5.73, p = 0.0000, \beta_{std} = +0.3106 SD; stripping archive mirrors flips the raw paywall link gap from +0.52 to -0.46 percentage points). Yet second-stage `Conflict × Paywalled` moderation is null across all outcomes (+0.0052, p = 0.5540 on reciprocity; -0.0046, p = 0.3565 on nested replies; -0.0012, p = 0.8034 on evidence; K = 3 Holm p = 1.0000; `Negativity × Paywalled` on nested replies +0.0107, Holm p = 0.0727), with positive conflict reciprocity on both open-access (+0.0117, p = 0.0252) and paywalled domains (+0.0169, p = 0.0383$).

> ***Clarifying Two Numbers on Paywall Mirrors (Audit Note from Singapore External Audit):*** *To make sure readers never confuse the raw comparison with the regression-adjusted estimate:* - ***$5.38× is the Raw Mean Ratio:*** *Across our 58 open-access outlets, an average of* ***0.202%*** *(0.002023) of comments contain an archive-mirror link (such as `archive.ph` or the Wayback Machine, posted for preservation). Across our 27 paywalled outlets, an average of* ***1.089%*** *(0.010892) of comments contain an archive-mirror link. Dividing the two raw averages gives 0.010892 / 0.002023 = 5.384× (or a raw difference of +0.887$ percentage points).* - ***+0.0092 (+0.92 percentage points, t = +5.73) is the Regression-Adjusted Slope:*** *When we control for topic beat, calendar quarter, time-of-day slot, comment volume splines, and headline features in Model 4, being on a paywalled domain increases the archive-mirror link rate by* ***+0.0092 (+0.92 percentage points, SE = 0.0016, t = +5.73, p < 0.0001, or +0.3106 SD)****. Both the raw $5.38× ratio and the controlled +0.92 percentage-point regression jump tell the exact same story: in an average 107$-comment thread on a paywalled newspaper, readers reliably post roughly one full-text archive mirror link so the whole community can read the article.*

Across all ten falsification gates (detailed in Appendix B), core conclusions hold: - **Per-Judge and Leave-One-Judge-Out Checks (Gate 1 & Check E):** Individual models confirm Stage-2 conflict reciprocity for Llama (+0.0117, t = +2.92), DeepSeek (+0.0144, t = +2.66), and Gemma (+0.0164, t = +3.53), while Qwen loads

slightly more on negativity (t = +2.50 vs. +1.32); all four three-judge leave-one-judge-out ensembles remain significant for conflict (+0.0114 to +0.0157, p ≤ 0.0162) and null for negativity (p ≥ 0.1787). Furthermore, the Singapore sister-lab external audit independently re-estimated Model 2 using **L1 median (quantile) regression** and confirmed that Moral Conflict's reciprocity effect remains highly significant at the median (+0.0108, t = +3.01, p = 0.0026), proving that our findings are not driven by extreme mean outliers. - **Unedited Newsroom URL Slugs (Gate 2):** Within lexical-slug outlets (excluding opaque path identifiers on *BBC* and *Nature*), stories with ≥ 0.60 slug-to-title overlap ($2167/3172 = 68.32% Stage 1; 1182/1797 = 65.78% Stage 2) yield +0.0140 on conflict reciprocity (t = +2.40, p = 0.0189) and +0.0001 on negativity (t = +0.03$), ruling out post-submission title edits. - **Selection, Splines, Windows, and Subsamples (Gates 3–9):** Heckman (1979) Inverse Mills Ratio adjustment (λ = +0.0228, t = +0.68; conflict +0.0137, t = +2.81), breakthrough inverse-probability weights (+0.0179, t = +3.47, p = 0.0008; combined weights +0.0202, t = +3.33), 20-vigintile volume fixed effects (+0.0135, t = +2.75), unadjusted volume (+0.0130, t = +2.70), 24-hour (+0.0129, t = +2.82) and 14-day (+0.0135, t = +2.98) horizons, no-lexicon specs (+0.0143, t = +3.28), $1%/99% winsorization (+0.0133, t = +2.88), and every top-10 outlet drop (t ∈ [+2.54, +2.90]) and single-year drop (t ∈ [+2.02, +2.84]; post-2023 interactions p ≥ 0.5219$) preserve inference.

Chronological OOS Task (2021–2023 \to 2024–2025)	**Constant RMSE**	**Controls R^2_{OOS}**	**Lexical R^2_{OOS}**	**Shrunk 4-Judge R^2_{OOS}**	**DM t vs. Ctrl (p)**	**DM t vs. Lex (p)**
Stage-1 Log Comments (N_{test} = 1422)	$1.3291	+0.8863%	+1.0654%	+1.6528%	-3.74 (0.0003)	-3.39 (0.0010$)
Stage-2 Reciprocity (With Splines) (N_{test} = 759)	$0.1076	+0.4129%	+0.4995%	+1.0461%	-2.33 (0.0217)	-2.15 (0.0342$)
Stage-2 Reciprocity (Ex-Ante t = 0) (N_{test} = 759)	$0.1076	+0.4086%	+0.4997%	+1.0542%	-2.35 (0.0206)	-2.15 (0.0335$)
Stage-2 Nested Share (Ex-Ante t = 0) (N_{test} = 759)	$0.1090	+1.6946%	+1.8861%	+2.2984%	-3.21 (0.0018)	-2.61 (0.0104$)
Stage-2 Evidence Rate (Ex-Ante t = 0) (N_{test} = 759)	$0.0678	+3.5737%	+3.5924%	+4.2182%	-2.48 (0.0149)	-2.59 (0.0111$)

Table 5 reports strict chronological out-of-sample evaluation (2021–2023 training, N = 2231 and N = 1355; 2024–2025 test, N = 1422 and N = 759; Diebold & Mariano, 2002). While unshrunk high-dimensional ordinary least squares overfits (R^2_{OOS} = -4.3383% and -5.9442%), ridge-shrunk four-judge semantic predictions beat

both structural controls and zero-LLM dictionaries across Stage-1 log comments and strictly ex-ante (t = 0, without realized volume splines) Stage-2 reciprocity, nested share, and documentary evidence ($p \leq 0.0335$).

> ***Plain-Language Method Box 4: What Are Chronological Out-of-Sample (R^2_{OOS}) Forecasts and the Diebold–Mariano (DM) Test? - Why Test on Future Years (2024–2025)?*** *In many social science papers, models are judged only by how well they fit the historical dataset they were estimated on ("in-sample fit"). A stricter test of whether a measurement is real is* ***chronological out-of-sample forecasting****: we freeze our statistical model using only stories from* ***2021–2023****, and then ask it to predict the conversation structure of unseen stories published in* ***2024–2025*** *($1422 Stage-1 stories and 759 Stage-2 trees)—using* ***only information visible at the exact second the link was posted (t = 0$)*** *(`Ex-Ante t = 0`). - **What Does R^2_{OOS} of +1.05% to +4.22% Mean?** Out-of-sample R^2_{OOS} measures the percentage reduction in squared prediction error on future 2024–2025 stories compared to guessing the historical 2021–2023 average. Notice that if you throw $85 outlet indicators and 28 time-slot indicators into a standard unpenalized regression (`unshrunk OLS`), the model memorizes 2021–2023 noise and performs worse than a simple average on 2024–2025 data (R^2_{OOS} = -4.3383% and -5.9442%). When properly regularized (`ridge-shrunk`), structural metadata and word-list dictionaries explain +0.41% to +3.59% of future conversation variance, and* ***adding our four-judge AI semantic framing scores raises out-of-sample predictive accuracy to +1.0542% on reciprocity, +2.2984% on nested reply share, and +4.2182%$ on external evidence links****. - **What Is the Diebold–Mariano (DM) Test?** Named after economists Francis Diebold and Roberto Mariano (2002), the DM t-statistic tests whether one forecasting model's prediction errors on the 2024–2025 holdout set are systematically smaller than a rival model's errors (clustered by test week). Here, negative DM t-statistics (-2.15 to -3.74, $p \leq 0.0342$) prove that our four-judge AI framing scores provide a* ***statistically significant improvement in real-world forecasting over both structural metadata and traditional word-counting dictionaries****—while the modest +1% to +4% magnitude reminds us that a 12-word headline is one meaningful signal alongside the full article text and breaking news events.*

---

5. Theoretical Stress-Testing and a Revised Public-Sphere Framework

Because our pre-registered tests falsified simple binary readings of H_1, H_2, and H_0, we conducted five post-data adversarial stress tests on the 226317 raw comments to construct a revised public-sphere framework (Table 6).

5.1 Agonistic Dyadic Closure: Trench Duels Versus Habermasian Reciprocity

Standard computational schemes interpret A \to B \to A reciprocity as Habermasian dialogic engagement (Graham & Witschge, 2003; Stromer-Galley, 2007). Why, then, did Moral-Conflict Framing raise reciprocity (+0.0135) while lowering speaker entropy (-0.0029) and documentary links (-0.0073)? Decomposing Total

Dyadic Reciprocity ($0.211650, or 21.16%) into an exact additive identity—**Unique-Dyad First Reciprocity Rate** (0.145407, or 14.54%) plus **Repeat-Dyad Ping-Pong Rate** (0.066243, or 6.62%, 31.30%$ of baseline reciprocity)—resolves the paradox.

> ***Plain-Language Method Box 5: How the Exact Additive Decomposition of A \to B \to A Reciprocity Works*** *Suppose a discussion thread has $100 person-to-person replies (* `Depth >= 2` *), and 21 of those replies complete an A \to B \to A turn (* `Total Dyadic Reciprocity Rate = 21.0%` *). How can we tell whether those 21$ reciprocal turns represent* ***21 different pairs of citizens listening and responding to each other once*** *(a broad town-hall conversation) versus* ***just 3 pairs of entrenched combatants arguing back and forth 7 times each*** *(a locked-in trench duel)?*
>
> *We split every single A \to B \to A turn in a thread into two mutually exclusive buckets: 1.* ***Unique-Dyad First Reciprocity:*** *The very first time an unordered pair of citizens {A, B} completes a three-turn A \to B \to A exchange in that thread. This measures* ***how widely back-and-forth dialogue spreads across distinct pairs of citizens*** *(averaging $14.54%$ of interpersonal replies). 2.* ***Repeat-Dyad Ping-Pong:*** *The second, third, fourth, or tenth time that* ***exact same pair {A, B}*** *trades another back-and-forth blow in the same thread (A \to B \to A \to B \to A \dots, averaging $6.62%$ of interpersonal replies).*
>
> *Because every reciprocal turn is either a pair's first exchange or a repeat volley within that same pair, the two rates add up to Total Reciprocity on every single row ($14.5407% + 6.6243% = 21.1650%$). And by a mathematical property of linear regression, when we run the exact same regression on the two pieces,* ***their regression slopes must also add up exactly to the total slope****: -* ***Total Reciprocity Slope (+0.013542, t = +2.82)*** *=* ***First-Time Unique Pairs (+0.002649, or $19.56%, t = +1.05, not significant)*** *+* ***Repeat Two-Person Ping-Pong (+0.010893, or 80.44%, t = +3.13, p = 0.0024$)****.*
>
> *This exact arithmetic decomposition is the smoking gun of our study:* ***four-fifths ($80.44%$) of Moral Conflict's apparent gain in "reciprocity" is just already-fighting pairs trading repeat ping-pong volleys, while dialogue among new pairs of citizens does not significantly expand at all.***

Exactly **$80.44% (+0.0109, SE = 0.0035, t = +3.13, p = 0.0024) of Moral Conflict's +0.0135 reciprocity gain comes from repeat back-and-forth volleys within already-fighting pairs**, whereas **first-time unique-dyad reciprocity is flat (+0.0026, SE = 0.0025, t = +1.05, p = 0.2950; 19.56%)**! Holding \log(V_{48}) splines fixed, Moral Conflict compresses Unique Speakers Per Comment by -0.0174 (SE = 0.0052, t = -3.34, p = 0.0013, off a 67.84% mean) and elevates Top-2 Speaker Share by +0.0074 (SE = 0.0031, t = +2.38, p = 0.0198, off 17.34%), explaining why dyads per speaker rise (+0.0130, t = +2.85, p = 0.0055) as the speaker pool contracts. Moreover, documentary erosion is not merely compositional: although A \to B \to A turns cite fewer links than non-reciprocal comments (7.77% versus 10.28%), Moral Conflict **directly depresses external citing inside non-reciprocal comments (-0.0079, SE = 0.0026, t = -3.00, p = 0.0035; within-reciprocity -0.0045, t = -0.84$)**. Both Habermasian deliberation (Habermas, 2006) and agonistic pluralism (Papacharissi, 2004; Karppinen, Moe, & Svensson, 2008; Rossini, 2020) must separate *interactive persistence (heat)* from

polycentric breadth and documentary grounding (light): moral-conflict headlines catalyze *agonistic dyadic closure*—protracted two-person trench duels with fewer facts.

5.2 The Two-Stage Gatekeeping Chokepoint

Estimating Model 1 across progressive comment hurdles (N = 3653) explains why Stage-1 upvotes rose while ≥ 15-comment breakthroughs did not. At **zero-to-one comment ignition** (≥ 1 comment, $43.25% base rate), both Negativity (+0.0306, t = +2.54, p = 0.0129; standalone +0.0495, t = +5.47) and Moral Conflict (+0.0300, t = +2.35, p = 0.0211; standalone +0.0526, t = +5.41) spark initial reactions. Yet effects fade by ≥ 5 comments (16.37%: conflict +0.0185, t = +1.84, p = 0.0700) and vanish across ≥ 10 (12.35%), ≥ 15 (10.90%), ≥ 30 (8.60%), ≥ 50 (6.60%), and ≥ 100 comments (3.78%, all |t| ≤ 1.62), with conditional \log(comments) \mid comments ≥ 1 (N = 1579) flat at -0.0078 (t = -0.11, reflecting newly ignited 1–4 comment threads alongside ≥ 150-comment megathreads at 2.41%, t = +1.80$). Individual psychophysiological capture (Soroka, Fournier, & Nir, 2019; Robertson et al., 2023) governs first contact, but non-personalized community curation (Wright & Street, 2007; Milli et al., 2025; Piccardi et al., 2025) acts as an institutional circuit-breaker against outrage breakthroughs.

5.3 Propositional Contestability Over Arbitrary Valence Cutoffs

Why was the pre-registered $0.50 four-cell contrast C_2 - C_1 non-significant on reciprocity (+0.0111, t = +0.96, p = 0.3389; grid check p = 0.3382) despite continuous Negativity equaling zero? Grid diagnostics across {0.35, 0.40, 0.50, 0.60} and sample medians (0.325 negativity, 0.20 conflict) reveal *sub-threshold conflict contamination*: because journalistic conflict is right-skewed (median 0.20), high cutoffs sweep moderate-blame stories into Pure Threat (41.08% of C_1 at 0.50 and 48.03% at 0.60 have C_i ≥ 0.30, versus 15.27% at 0.35). Where C_1 cleanly isolates low-blame adversity—at 0.35 (C_1 = +0.0029, p = 0.6476 vs. C_2 = +0.0302, p = 0.0001), 0.40, and the median—C_2 - C_1 is significant (+0.0273, p = 0.0023; +0.0187, p = 0.0385; +0.0199, p = 0.0437). Crucially, **Non-Negative Debate (C_3: N_i < 0.50, C_i ≥ 0.50, n = 189)**—policy or legal disagreement without tragedy valence—significantly boosts reciprocity (+0.0212, p = 0.0302; +0.0254, p = 0.0060 at 0.60), nested share (+0.0179, p = 0.0083), and 90th-percentile depth (+0.2694, p = 0.0037$). The active mechanism is **propositional contestability and responsibility attribution** (Entman, 1993; Semetko & Valkenburg, 2000; Shi et al., 2019), not negative tone.

Theoretical Clash	Key Empirical Diagnostic	Revised Public-Sphere Proposition
1. Trench Duels vs. Reciprocity	$80.44% (+0.0109, t = +3.13) of +0.0135 reciprocity is repeat dyad ping-pong; speakers/comment -0.0174 (t = -3.34); non-ABA evidence -0.0079 (t = -3.00$).	**Agonistic Dyadic Closure**: Moral conflict drives locked-in two-person duels with narrower participation and fewer documentary links.
		Curation Circuit-Breaker: Affective framing sparks individual clicks/first

Theoretical Clash	Key Empirical Diagnostic	Revised Public-Sphere Proposition
2. Two-Stage Chokepoint	$0 \to 1 ignition positive (+0.0306, t = +2.54; +0.0300, t = +2.35); ≥ 15 breakthrough null (t = +1.52, -0.51$).	comments, but non-algorithmic curation blocks front-page breakthroughs.
3. Contestability vs. Valence	Continuous negativity $0.00 on five of five outcomes; C_2 - C_1 +0.0273 (p = 0.0023) at 0.35 vs. null at 0.50; Non-Negative Debate C_3 +0.0212 (p = 0.0302$).	**Propositional Contestability**: Competing actors and responsibility attribution—not sad or threatening valence—drive counter-arguing.
4. Journalistic Field Scope	Civic beats (N = 962): reciprocity +0.0165 (t = +2.43); Specialist beats (N = 1151): nested +0.0112 (t = +2.30), evidence -0.0066 (t = -2.18); interaction p = 0.4173.	**Field Precision With Broad Link Erosion**: Reciprocity precision is sharpest in civic reporting, while documentary link erosion spans beats.
5. Networked Paywall Repair	Archive mirrors jump $5.38× (+0.0092, t = +5.73), neutralizing H_4 (Holm p = 1.0000); ex-ante t = 0 OOS R^2 +1.0542% to +4.2182%$.	**Networked Epistemic Repair**: Communal mirror links equalize text access in link-permissive forums; semantic framing beats word lists OOS.

5.4 Journalistic Field Scope and Networked Paywall Repair

Across journalistic fields (Benson, 2006), reciprocity precision is strongest in National/International Wire (N = 831, M = 27: +0.0199, t = +2.76, p = 0.0105) and combined Civic/Investigative reporting (N = 962, M = 41: +0.0165, t = +2.43, p = 0.0195), whereas Technology-in-Society (N = 458, M = 11) is imprecise (-0.0031, t = -0.26, $95% CI [-0.0300, +0.0237]). Respecting interaction discipline, full-panel beat interactions are non-significant (`Conflict × Civic` +0.0066, p = 0.4173; `Conflict × Tech` -0.0079, p = 0.3599), and Specialist Science, Technology, and Culture beats (N = 1151, M = 44) still exhibit significant nested-reply expansion (+0.0112, p = 0.0267) and evidence-link erosion (-0.0066, p = 0.0320 in the full-panel interaction model). Meanwhile, the 5.38$-fold surge in communal archive mirrors on paywalled outlets demonstrates *networked epistemic repair* (Barzilai-Nahon, 2008; Carlson, 2015), equalizing full-text access for downstream commenters.

6. Limitations, Scope Conditions, and Conclusion

Four scope conditions bound our conclusions. First, despite four-way outlet, quarter, slot, and beat fixed effects and replication on N = 1182 verbatim URL-slug headlines, observational headlines reflect both underlying events and editorial choices; all findings represent conditional structural associations rather than randomized causal effects. Second, Hacker News is a high-literacy, anti-insult-moderated threaded forum without

personalized outrage feeds; platforms that throttle outbound links (blocking archive-mirror repair) or algorithmically reward abuse may exhibit steeper crowding-out (Milli et al., 2025; Piccardi et al., 2025). Third, professional journalism displays substantial conflict zero-mass ($31.37% exact zero and 43.33% ≤ 0.05 in Stage 1; 23.37% and 34.30% in Stage 2), and two-part hurdle checks show reciprocity shifts strongly across the extensive onset of conflict (`Conflict > 0.05` +0.0302, t = +3.39, p = 0.0011, vs. +0.0074, t = +1.34 conditional intensity). Fourth, modest submission-time out-of-sample R^2 (+1.0542% to +4.2182%$) confirms that headline framing is a genuine but bounded influence alongside full-article substance.

Ultimately, digital engagement is not one thing. Purely negative news sparks fleeting individual clicks without altering public dialogue. Moral-conflict and blame headlines reshape public conversation—not into silent monologues or inclusive Habermasian inquiry, but into deep, concentrated, documentary-light trench duels. Evaluating the health of digital public spheres requires measuring not merely whether citizens reply, but how broadly participation circulates and whether disagreement remains anchored in shared documentary evidence.

Plain-Language Glossary of Concepts and Methods

Term	Plain-Language Meaning in This Paper
Agonistic Dyadic Closure ("Trench Duels")	Our core theoretical concept for what happens under moral-conflict headlines: comments become deeper and show more back-and-forth replies, *not* because a broader community is deliberating together, but because a few pairs of commenters lock into repeat two-person ping-pong arguments while fewer unique citizens speak and fewer external facts are cited.
Collider / Selection Bias & Two-Stage Hurdle Design	If you only study threads that became popular (≥ 15 comments), you miss all the stories that died with zero or two comments. Our **two-stage hurdle design** models (1) which submissions cross the discussion hurdle in Stage 1 (N = 3653) and (2) how conversations unfold inside active threads in Stage 2 (N = 2114).
Diebold–Mariano (DM) Test & Out-of-Sample R^2_{OOS}	A fair forecasting tournament where models are trained only on 2021–2023 stories and tested on unseen 2024–2025 stories. R^2_{OOS} measures the percentage reduction in prediction error compared to guessing the historical average; the DM t-test verifies whether our AI framing scores predict future thread geometry significantly better than metadata and word lists alone.
Dyadic Reciprocity (A \to B \to A) & Additive Ping-Pong Decomposition	A three-turn conversation loop where Citizen A posts, Citizen B replies to A, and Citizen A replies back to B. We split every A \to B \to A turn into **Unique-Dyad First Reciprocity** (the very first time pair {A,B} completes a loop) plus **Repeat-Dyad Ping-Pong** (every subsequent loop between that exact same pair {A,B}).
Fixed Effects (Four-Way)	A statistical control that subtracts out the baseline average of every news outlet ($85), every calendar quarter (20), every day-of-week-by-6-hour

Term	Plain-Language Meaning in This Paper
	posting window (28$), and every URL topic beat so that comparisons happen strictly *within* the same outlet, beat, and time window.
Headline Negativity vs. Moral-Conflict Framing	**Negativity** measures sad, threatening, or crisis tone (which does not require a human villain, such as a natural disaster or disease). **Moral-Conflict Framing** measures blame, accusation, wrongdoing, or two sides clashing over values or policy.
Holm Multiplicity Correction	A strict statistical penalty applied when testing multiple outcomes at once (such as our K = 4 primary slopes or K = 5 Stage-2 tree metrics) so that chance fluctuations cannot masquerade as significant discoveries.
Intraclass Correlation (ICC) & Joint Evaluation (JEV)	A $0.00-to-1.00 measure of inter-rater agreement across our four independent open-weight AI language models (Llama 3.3 70B, Qwen 2.5 72B, DeepSeek V3, and Gemma 3 27B). Our four-judge average (**JEV**) achieves ICC(2,4) > 0.96$.
Inverse-Probability Weights (IPW)	Survey-style weights recording how many stories in the full 220696-story universe each stratified sample row represents. Used to verify that our unweighted within-outlet regression models generalize to the full 5-year universe.
Networked Epistemic Repair	How online readers collectively overcome a news paywall: when an article is locked behind a subscription wall, commenters post $5.38 times as many full-text archive-mirror links (1.089% vs. 0.202%$), restoring shared access to the article text for downstream commenters.
Restricted Cubic Splines of Comment Volume	A flexible curve controlling for total thread size ($\log V_{48}$) across four knot points (≈ 16, 39, 94, and 342 comments) so that natural differences between small and large threads are never mistaken for headline framing effects.
Speaker Shannon Entropy	An information-theory measure of speaker equality from $0.00 (one voice monopolizes the thread) to 1.00$ (every participant speaks an equal number of times).
Two-Way Clustered Standard Errors	Error bars adjusted simultaneously for correlation across stories from the same news outlet (G = 85) and stories published during the same calendar week (G = 262 or 261).

Appendix A: Formal Econometric Specifications and Estimation Details

For quantitative social scientists and methodologists, this appendix records the exact mathematical equations estimated in Section 4.

A.1 Stage-1 Unconditional Crowd Gatekeeping Specification (Model 1)

On the Stage-1 stratified panel (N = 3653 submissions across G = 85 outlets and G = 262 calendar weeks), for each Stage-1 outcome $Y_{1,i}$ (Discussed Breakthrough ≥ 15 comments, Log Upvote Points, and Log Comment Volume), we estimate the four-way fixed-effects linear model:

$$Y_{1,i} = \beta_{1,neg} N_{i,z} + \beta_{1,conf} C_{i,z} + X_i'\boldsymbol{\theta}_1 + \alpha_{d(i)} + \tau_{q(i)} + \omega_{s(i)} + \delta_{b(i)} + \varepsilon_{1,i}$$

where $N_{i,z}$ and $C_{i,z}$ are the standardized (z-scored, mean $0, SD 1) four-judge consensus scores for Headline Negativity and Moral-Conflict and Blame Framing; X_i contains standardized structural and lexical headline controls (character length, word count, question mark indicator, colon/dash punctuation indicator, digits/ numbers indicator, direct quote indicator, and deterministic word-list dictionary rates across 2487 negative words and 199 conflict words); and \alpha_{d(i)}, \tau_{q(i)}, \omega_{s(i)}, \delta_{b(i)} denote absorbed fixed effects for outlet domain (85), year-quarter (20), UTC day-of-week-by-6-hour submission slot (28$), and URL section beat.

A.2 Stage-2 Size-Adjusted Conversation Geometry Specification (Models 2, 3, and 4)

On the Stage-2 deliberative tree panel (N = 2114 recursive 48-hour comment trees across G = 85 outlets and G = 261 weeks), for each deliberative outcome $Y_{2,i}$, we estimate:

$$Y_{2,i} = \beta_{2,neg} N_{i,z} + \beta_{2,conf} C_{i,z} + \sum_{k=1}^{3} \phi_k B_k(\log V_{i,48}) + X_i'\boldsymbol{\theta}_2 + \alpha_{d(i)} + \tau_{q(i)} + \omega_{s(i)} + \delta_{b(i)} + \varepsilon_{2,i}$$

where $B_1(\log V_{i,48})$, $B_2(\log V_{i,48})$, $B_3(\log V_{i,48})$ are the three Harrell restricted cubic spline basis functions of log 48-hour valid comment volume $\log(V_{i,48})$ evaluated at the pre-registered 5th, 35th, 65th, and 95th percentile knots [2.772589, 3.663562, 4.543295, 5.834811] (corresponding to approximately [16.0, 39.0, 94.0, 342.0] valid 48-hour comments). Model 3 tests the continuous Wald linear contrast H_3: $\beta_{2,conf} - \beta_{2,neg} = 0$ and the four-cell framing typology (C_0 Neutral/Constructive reference vs. C_1 Pure Threat/Tragedy, C_2 Moral-Conflict/Blame, and C_3 Non-Negative Debate). Model 4 adds interactions with the binary outlet paywall indicator $Paywalled_{d(i)} \in \{0, 1\}$ (M = 27 paywalled vs. M = 58 open-access domains; when testing the direct main effect of $Paywalled_{d(i)}$ on archive-mirror link rates in stage one of Model 4, domain fixed effects $\alpha_{d(i)}$ are replaced by curated outlet-beat fixed effects because paywall status is invariant within domain).

A.3 Two-Way Clustered Covariance Matrix (Cameron, Gelbach, & Miller, 2011)

Across all linear models, variance-covariance matrices are computed via the Cameron, Gelbach, and Miller (2011) multi-way cluster-robust estimator on the within-transformed residuals after iterative demeaning across all fixed-effect dimensions:

$$\widehat{V}_{CGM} = \widehat{V}_{domain} + \widehat{V}_{week} - \widehat{V}_{domain\capweek}$$

with small-sample degrees-of-freedom correction and t-distribution inference evaluated at df = \min(G_{domain}, G_{week}) - 1 = 84. Any negative eigenvalues in the finite-sample intersection subtraction are floored at zero via spectral projection. Our custom sparse solver is verified at runtime against standard statistical packages to machine precision ($< 10^{-15}$ discrepancy).

Appendix B: Complete Summary of the Ten Pre-Registered Falsification Gates (Gates 1–10)

Gate	Methodological Threat Tested	Empirical Specification & Diagnostic Result	Verdict
Gate 1	Single-LLM bias or rater noise	Re-computed textbook Shrout–Fleiss ICCs from raw ballots (ICC(2,4) = 0.9665, 0.9650 in Stage 1; $0.9638, 0.9623 in Stage 2). Standalone regressions confirm Stage-2 conflict reciprocity on Llama (+0.0117, t = +2.92), DeepSeek (+0.0144, t = +2.66), and Gemma (+0.0164, t = +3.53$).	**PASS**
Gate 2	Post-submission headline editing by users or moderators	Restricted to stories with lexical URL slugs and ≥ 0.60 token overlap between URL slug and submitted title ($2167/3172 = 68.32% Stage 1; 1182/1797 = 65.78% Stage 2). Stage-2 conflict reciprocity holds at +0.0140 (t = +2.40, p = 0.0189); negativity remains zero (+0.0001, t = +0.03$).	**PASS**
Gate 3	Stage-1 to Stage-2 viral-survival selection bias	Estimated Stage-1 Probit selection equation (pseudo-R^2 = 0.0436) and inserted Heckman (1979) Inverse Mills Ratio (λ = +0.0228, t = +0.68, p = 0.4975): conflict reciprocity unchanged at +0.0137 (t = +2.81). Inverse-breakthrough-probability weighting strengthens conflict reciprocity to +0.0179 (t = +3.47, p = 0.0008), and combined universe-plus-breakthrough IPW yields +0.0202 (t = +3.33, p = 0.0013).	**PASS**
Gate 4	Functional form of comment-volume control	Replaced 4-knot cubic splines of \log(V_{48}) with 20 non-parametric vigintile volume fixed effects (+0.0135, t = +2.75) and with zero volume control (+0.0130, t = +2.70). All four key outcomes remain invariant.	**PASS**
Gate 5	Arbitrary 48-hour observation cutoff	Re-estimated across strict 24-hour trees (N = 1996: conflict reciprocity +0.0129, t = +2.82, p = 0.0060) and full 14-day conversation lifecycles (N = 2114: conflict reciprocity +0.0135, t = +2.98, p = 0.0038; nested share +0.0122, t = +3.57, p = 0.0006).	**PASS**
Gate 6		Excluded $562 A \to A self-replies (0.25% of 48h comments; mean 0.28%). Separated archive-mirror links (1.089%	**PASS**

Gate	Methodological Threat Tested	Empirical Specification & Diagnostic Result	Verdict
	Author self-reply monologues (A \to A) & archive mirror links	paywalled vs. 0.202% open-access, 5.38× ratio; regression slope +0.0092, t = +5.73) from genuine external documentary links (stripping mirrors flips the paywall link difference from +0.52 to -0.46$ percentage points).	
Gate 7	Survey sampling weights & nonlinear count/binary models	Stage-1 IPW universe-weighted breakthrough remains null (t = +0.21 and +0.66); Stage-1 Logit breakthrough remains null (z = +1.52 and +0.10); Stage-1 Poisson PPML confirms positive Moral-Conflict slope on total comment counts (+0.2869, z = +2.16, p = 0.0311).	**PASS**
Gate 8	Over-partialling by lexical dictionaries & extreme outliers	Dropping word-list dictionary controls yields conflict reciprocity +0.0143 (t = +3.28, p = 0.0015) and negativity +0.0014 (t = +0.44). Sample-median 4-cell contrast (C_2 - C_1) is significant at +0.0199 (t = +2.05, p = 0.0437). Winsorizing continuous outcomes at $1%/99% yields conflict reciprocity +0.0133 (t = +2.88, p = 0.0051$).	**PASS**
Gate 9	Dominance by a single outlet, beat, or calendar year	Dropping each top-10 outlet one at a time preserves conflict reciprocity (b ∈ [+0.0114, +0.0144], t ∈ [+2.54, +2.90]). Dropping each calendar year 2021–2025 one at a time preserves conflict reciprocity (b ∈ [+0.0106, +0.0156], t ∈ [+2.02, +2.84]). Post-2023 regime interactions are null (p = 0.5219 and p = 0.7472).	**PASS**
Gate 10 & Check F	In-sample overfitting vs. true out-of-sample generalization	Strict chronological split (Train 2021–2023 \to Test 2024–2025). Ridge-shrunk 4-judge semantic predictions significantly beat both structural controls and word-list dictionaries across Stage-1 log comments (R^2_{OOS} = +1.6528%, DM\ p ≤ 0.0010) and strictly ex-ante (t = 0) Stage-2 reciprocity (+1.0542%, DM\ p ≤ 0.0335), nested share (+2.2984%, DM\ p ≤ 0.0104), and evidence links (+4.2182%, DM\ p ≤ 0.0149).	**PASS**

Appendix C: Independent Verification & Audit Provenance Ledger (Boston Self-Audit & Singapore Sister-Lab External Audit)

To ensure complete transparency for interdisciplinary readers and peer reviewers, Table C1 summarizes the two independent forensic audits completed on November 19, 2026 prior to the release of Reader Edition v1.1:

Audit Dimension	Boston Internal Forensic Self-Audit (Nov 19, 2026)	Singapore Sister-Lab Red-Team Audit (Audit 37, DeepSeek V4.1 Flash)	Integrated Resolution in Edition v1.1
Data Authenticity Verdict	**Real Archival Data** ($0 synthetic, simulated, or fallback rows across 3653 Stage-1 stories, 2114 Stage-2 trees, and 21592 live OpenRouter ballots).	**Real & Replicated** (20/20 random story IDs matched live HN Algolia API exactly on title, URL, date, and integer points; live comment trees and 21592$ ballots verified).	Full pipeline lineage and raw cache hashes documented in the standalone self-audit report.
Econometric & Hash Reproduction	$19/19 artifact SHA-256 hashes and 465-line verification log reproduced byte-for-byte (0 violations).	14/14 bundle hashes and every tested regression coefficient, SE, t-stat, Holm p-value, and 80.44% / 19.56%$ ping-pong split reproduced to 6 decimal places.	Added Appendix A and Method Boxes 3–5 explaining the Harrell 4-knot spline basis and additive ping-pong identity.
Alternative Estimator Checks	Verified Logit, Poisson PPML, 20-vigintile volume FEs, Heckman IMR, IPW weights, and $1%/99% winsorization.	Independently ran **L1 median (quantile) regression** on Stage-2 reciprocity: Moral Conflict remained highly significant (+0.0108, t = +3.01, p = 0.0026) while Negativity remained null (+0.0062, t = +1.82).	Reported in Section 4.3 alongside leave-one-judge-out checks (t ∈ [+2.45, +3.22]$).
Clarification 1: Unweighted Primary Models vs. IPW	Noted that primary multi-way FE models estimate unweighted within-outlet slopes while IPW weights are tested in Gates 3 & 7.	Flagged that readers might assume primary models are IPW-weighted unless explicitly explained.	Explained in **Plain-Language Method Box 1** (Section 3.1) and Appendix B (Gates 3 & 7).
Clarification 2: Raw $5.38× Paywall Ratio vs. +0.0092 Adjusted Slope	Verified both the raw ratio (1.089% / 0.202% = 5.384×) and regression slope (+0.0092, t = +5.73).	Noted that the abstract leads with the raw 5.38× ratio while Section 4.3 reports the regression-adjusted +0.0092$ slope.	Both figures now appear together in the Abstract and are explained side-by-side in **Section 4.3**.
Clarification 3: Raw Algolia Hit Count ($221497)	Reconciled an initial self-audit draft typo (238311) back to the true on-disk count N = 221497.	Verified 221497 \to 220696 \to 3653 / 2114$ waterfall directly from the replication bundle.	Explicitly documented in **Plain-Language Method Box 1** (Section 3.1).

References

- Barzilai-Nahon, K. (2008). Toward a theory of network gatekeeping: A framework for exploring information control. *Journal of the American Society for Information Science and Technology*, 59(9), 1493–1512. https://doi.org/10.1002/asi.20857
- Benson, R. (2006). News media as a "journalistic field": What Bourdieu adds to new institutionalism, and vice versa. *Political Communication*, 23(2), 187–202. https://doi.org/10.1080/10584600600629802
- Boczkowski, P. J., & Peer, L. (2011). The choice gap: The divergent online news preferences of journalists and consumers. *Journal of Communication*, 61(5), 857–876. https://doi.org/10.1111/j.1460-2466.2011.01582.x
- Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114
- Bright, J. (2016). The social news gap: How news reading and news sharing diverge. *Journal of Communication*, 66(3), 343–365. https://doi.org/10.1111/jcom.12232
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2), 238–249. https://doi.org/10.1198/jbes.2010.07136
- Carlson, M. (2015). Metajournalistic discourse and the meanings of journalism: Definitional control, boundary work, and legitimation. *Communication Theory*, 26(4), 349–368. https://doi.org/10.1111/comt.12088
- Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1), 134–144. https://doi.org/10.1198/073500102753410444
- Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x
- Frieß, D., & Eilders, C. (2015). A systematic review of online deliberation research. *Policy & Internet*, 7(3), 319–339. https://doi.org/10.1002/poi3.95
- Graham, T., & Witschge, T. (2003). In search of online deliberation: Towards a new method for examining the quality of online discussions. *Communications*, 28(2), 173–204. https://doi.org/10.1515/comm.2003.012
- Habermas, J. (2006). Political communication in media society: Does democracy still enjoy an epistemic dimension? *Communication Theory*, 16(4), 411–426. https://doi.org/10.1111/j.1468-2885.2006.00280.x
- Halpern, D., & Gibbs, J. (2012). Social media as a catalyst for online deliberation? *Computers in Human Behavior*, 29(3), 1159–1168. https://doi.org/10.1016/j.chb.2012.10.008
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161. https://doi.org/10.2307/1912352
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. https://doi.org/10.2307/4615733
- Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2018). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22, 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034
- Jaidka, K., Zhou, A., Lelkes, Y., Egelhofer, J., & Lecheler, S. (2021). Beyond anonymity: Network affordances, under deindividuation, improve social media discussion quality. *Journal of Computer-Mediated Communication*, 27(1), zmab019. https://doi.org/10.1093/jcmc/zmab019

- Karppinen, K., Moe, H., & Svensson, J. (2008). Habermas, Mouffe and political communication: A case for theoretical eclecticism. *Javnost - The Public*, 15(3), 5–21. https://doi.org/10.1080/13183222.2008.11008973
- Marcus, G. E., MacKuen, M., & Neuman, W. R. (2011). Parsimony and complexity: Developing and testing theories of affective intelligence. *Political Psychology*, 32(2), 323–336. https://doi.org/10.1111/j.1467-9221.2010.00806.x
- Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., & Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, 4(3), pgaf062. https://doi.org/10.1093/pnasnexus/pgaf062
- Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, 67(4), 586–609. https://doi.org/10.1111/jcom.12312
- Mutz, D. C. (2002). Cross-cutting social networks: Testing democratic theory in practice. *American Political Science Review*, 96(1), 111–126. https://doi.org/10.1017/s0003055402004264
- Myllylahti, M. (2013). Newspaper paywalls—the hype and the reality. *Digital Journalism*, 2(2), 179–194. https://doi.org/10.1080/21670811.2013.813214
- Neuman, W. R., Guggenheim, L., Jang, S. M., & Bae, S. Y. (2014). The dynamics of public attention: Agenda-setting theory meets big data. *Journal of Communication*, 64(2), 193–214. https://doi.org/10.1111/jcom.12088
- Noëlle-Neumann, E. (1974). The spiral of silence: A theory of public opinion. *Journal of Communication*, 24(2), 43–51. https://doi.org/10.1111/j.1460-2466.1974.tb00367.x
- Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, 6(2), 259–283. https://doi.org/10.1177/1461444804041444
- Piccardi, T., Saveski, M., Jia, C., Hancock, J., Tsai, J. L., & Bernstein, M. S. (2025). Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science*, 387(6732), eadu5584. https://doi.org/10.1126/science.adu5584
- Rathje, S., Van Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26), e2024292118. https://doi.org/10.1073/pnas.2024292118
- Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. *Nature Human Behaviour*, 7(5), 812–822. https://doi.org/10.1038/s41562-023-01538-4
- Rossini, P. (2020). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49(3), 399–425. https://doi.org/10.1177/0093650220921314
- Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of Communication*, 49(1), 103–122. https://doi.org/10.1111/j.1460-2466.1999.tb02784.x
- Semetko, H. A., & Valkenburg, P. M. (2000). Framing European politics: A content analysis of press and television news. *Journal of Communication*, 50(2), 93–109. https://doi.org/10.1111/j.1460-2466.2000.tb02843.x
- Shi, F., Teplitskiy, M., Duede, E., & Evans, J. A. (2019). The wisdom of polarized crowds. *Nature Human Behaviour*, 3(4), 329–336. https://doi.org/10.1038/s41562-019-0541-6
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420
- Sjøvaag, H. (2015). Introducing the paywall: A case study of content changes in three online newspapers. *Journalism Practice*, 10(3), 304–322. https://doi.org/10.1080/17512786.2015.1017595

- Soroka, S. (2014). *Negativity in democratic politics: Causes and consequences*. Cambridge University Press. https://doi.org/10.1017/cbo9781107477971
- Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, 116(38), 18888–18892. https://doi.org/10.1073/pnas.1908369116
- Soroka, S., & McAdams, S. (2015). News, politics, and negativity. *Political Communication*, 32(1), 1–22. https://doi.org/10.1080/10584609.2014.881942
- Stromer-Galley, J. (2007). Measuring deliberation's content: A coding scheme. *Journal of Deliberative Democracy*, 3(1), 12. https://doi.org/10.16997/jdd.50
- Trussler, M., & Soroka, S. (2014). Consumer demand for cynical and negative news frames. *The International Journal of Press/Politics*, 19(3), 360–379. https://doi.org/10.1177/1940161214524832
- Van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C., Matthes, J., Hopmann, D., Salgado, S., Reinemann, C., Papathanassopoulos, S., Maier, M., Sheafer, T., Stanyer, J., Štětka, V., & Theocharis, Y. (2021). Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association*, 41(1), 3–27. https://doi.org/10.1080/23808985.2017.1288551
- Wright, S., & Street, J. (2007). Democracy, deliberation and design: The case of online discussion forums. *New Media & Society*, 9(5), 849–869. https://doi.org/10.1177/1461444807081230