

Forewarned Is Not Forearmed: How Coordinated LLMs Shift Reader Opinions in News Comment Sections

Wooyoung Go¹ Yujin Potter² Hyounghick Kim³ Seungjin Ryu⁴ Yongdae Kim^{1†}

¹KAIST ²University of California, Berkeley ³Sungkyunkwan University ⁴NSR

gwy876@gmail.com, yujinyujin9393@berkeley.edu, hyoung@skku.edu, sjryu@nsr.re.kr, yongdaek@kaist.ac.kr

Abstract

Large language models (LLMs) can generate human-like online comments at scale, raising concerns that coordinated AI campaigns could manipulate public opinion. While prior work has shown that single AI messages can persuade individual readers, it remains untested whether coordinated multi-agent LLM campaigns, in which multiple agents jointly plan, divide roles, and manage visibility, can shift reader opinions in news comment sections, and whether per-comment AI labels can mitigate such influence. We address these questions through a controlled experiment with 1,000 U.S. adults reading news articles on a Reddit-like platform. Each participant was randomly assigned to one of six conditions, crossing three comment sources (human, independently generated AI, or coordinated multi-agent LLM) with two label settings (per-comment “Likely AI-generated” badges from a deployed detector, shown or hidden). Coordinated LLM comments did not increase average persuasion over independently generated AI comments, but disproportionately converted readers who initially disagreed and compressed opinion diversity by 12%, manufacturing apparent consensus where genuine disagreement existed. The labels reduced the perceived believability of comments but did not reliably reduce actual opinion change. These findings reveal what we call a *Knowing-Resisting dissociation* between source recognition and behavioral resistance, and indicate that per-comment disclosure alone is insufficient to counter coordinated LLM opinion campaigns.

1 Introduction

Online comment sections are a key venue where readers passively encounter many voices and form opinions on news. Large language models (LLMs) can generate comments at human-level quality, evading existing AI-text detectors at scale, raising the prospect that campaign operators could orchestrate

many LLM instances to flood these sections with content pushing a single narrative. [Schroeder et al. \(2026\)](#) warn of precisely such “AI Swarm” threats: a synthetic intervention into existing discourse that could manufacture the appearance of public consensus where genuine disagreement exists. The natural safeguard is to label AI-generated comments via deployed detectors so that readers can recognize and discount them. Yet whether such interventions actually shift reader opinions, and whether per-comment labels can mitigate them, remains empirically untested.

A growing body of research shows that AI-generated text can shift human opinions through single messages ([Bai et al., 2025](#); [Goldstein et al., 2024](#)) or interactive dialogues ([Potter et al., 2024](#); [Lai et al., 2025](#)). But news comment ecosystems differ structurally from these one-on-one paradigms, exposing passive readers to many voices alongside upvotes and popularity rankings that signal collective agreement. Yet no existing study addresses this setting. [Durmus et al. \(2024\)](#) evaluated AI arguments in single-message settings without social context; [Lin and Zhang \(2026\)](#) tested AI-labeled posts in a Weibo mockup and found a truth–falsity crossover, in which labels reduced credibility of accurate content while boosting that of misinformation; an unpublished Reddit field deployment reported higher AI persuasion rates but was withdrawn ([r/changemyview moderators, 2025](#)). None combines multiple AI voices, social-proof cues, pre-post opinion measurement, and a labeling countermeasure. Article 50 of the EU AI Act and emerging U.S. disclosure mandates assume that labeling AI content protects users, but this assumption remains untested in the very setting where such labels would be deployed.

We address this gap with a large-scale user study and a new multi-agent LLM pipeline. We conduct a controlled experiment with 1,000 U.S. adults on a Reddit-like platform under one of three comment-

†: Corresponding author

source conditions: organic Human comments, independently generated AI comments (LGCM: 11 LLMs producing comments in parallel without shared context), and coordinated multi-agent AI comments (MACM: a Campaign Planner designs the narrative arc and engagement roles for a unified 20-comment campaign; Appendix C). The LGCM-vs-MACM contrast isolates the effect of coordination from AI generation itself. Half of the participants see per-comment AI labels from a deployed detector (Go et al., 2025), enabling a direct test of the disclosure assumption. This design provides a first controlled test of AI-generated comments as a persuasion channel in a comment ecosystem with social-proof cues.

We organize the study around four research questions. (RQ1) Do AI-generated comments shift opinions in the campaign direction? (RQ2) Does multi-agent coordination amplify persuasion? (RQ3) Do coordinated AI comments reduce opinion diversity? (RQ4) Do per-comment AI labels mitigate opinion manipulation? Our results show that AI comments shift opinions, especially among those who initially disagreed. Under multi-agent coordination, those who initially disagreed are more strongly persuaded, and opinion diversity narrows. AI labels reduce perceived comment believability without reducing persuasion, a pattern we call the *Knowing-Resisting* dissociation.

We make three contributions. (i) MACM, a vertical multi-agent LLM coordination pipeline integrating centralized campaign planning, role-differentiated sequential generation, and ranking-aware visibility management (§3.2). (ii) A 3 (comment source) $\times$ 2 (AI label) between-subjects experiment with 1,000 U.S. adults providing the first controlled test of coordinated LLM persuasion in a news comment ecosystem with social-proof cues (§4). (iii) Evidence that coordinated AI converts dissenters, compresses opinion diversity by 12%, and survives per-comment disclosure, identifying a *Knowing-Resisting* dissociation between source recognition and behavioral resistance (§5).

2 Related Work

AI persuasion beyond one-on-one settings. Research on AI persuasion has mostly studied one-on-one settings (single human, single AI): single-message exposure (Bai et al., 2025; Goldstein et al., 2024) and interactive dialogues (Potter et al., 2024; Salvi et al., 2025; Lin et al., 2025). On the other

hand, the comment ecosystem we study is structurally different: a passive reader encounters many coordinated AI voices alongside upvotes, ranking, and signals of public opinion. Social-proof and ranking cues (Centola and Macy, 2007; Sundar, 2008; Noelle-Neumann, 1974) may amplify AI influence, yet cognitive overload may attenuate it; this uncertainty motivates our first question: whether AI comments can shift opinions in such an ecosystem. An unpublished Reddit field deployment illustrated practical reach but was withdrawn (r/changemyview moderators, 2025).

Multi-agent coordination and opinion diversity. Multi-agent persuasion research spans *internal collaboration* (Ma et al., 2025) and *external influence* (Song et al., 2025), but no study combines both with human readers in a comment ecosystem. Diversity is a prerequisite for collective intelligence (Surowiecki, 2004), and both human-to-human (Lorenz et al., 2011) and LLM-to-LLM (Chuang et al., 2024) influence reduce it. Whether coordinated AI-to-human influence destroys *human* opinion diversity, the “Undermined Collective Intelligence” pathway warned by Schroeder et al. (2026), remains untested, motivating our questions about whether coordination amplifies persuasion and whether it narrows opinion diversity.

AI content labeling. The Elaboration Likelihood Model (Petty and Cacioppo, 1986) predicts that labels may lower source credibility (peripheral route) without blocking argument influence (central route), a possible perception–behavior dissociation. Empirical findings diverge: Lin and Zhang (2026) report a *truth–falsity crossover* in a Weibo mockup, Gallegos et al. (2026) find labels do not reduce persuasion, and Lai et al. (2025) show that lower trust in AI does not necessarily prevent behavioral influence. No study measures both the mitigation benefit and side effects of AI labeling in a high-social-proof ecosystem; this gap motivates our question about whether per-comment AI labels can mitigate such manipulation.

We provide the first controlled study combining these elements, measuring whether AI-generated comments shift opinions, whether multi-agent coordination amplifies that effect, and whether per-comment AI labels can mitigate it.

3 Comment Generation Strategies

We design two AI comment-generation pipelines and compare them against authentic human comments from Reddit r/news. Both AI pipelines target a fixed campaign direction (promote or suppress) but differ in whether comments are generated independently or as a coordinated campaign. In this work, we define *coordination* as a campaign-level process in which multiple agents jointly pursue a shared manipulation objective through centralized planning, role assignment, contextual awareness, and visibility management. MACM implements these mechanisms; LGCM removes them and generates comments independently.

3.1 LGCM: Uncoordinated Baseline

LGCM (LLM-Generated Comment Manipulation) adapts the LGC framework (Go et al., 2025) (Figure 1): for each article, 11 frontier LLMs across three major providers (OpenAI / Anthropic / Google), representing commercially accessible models and producing varied writing styles that simulate a multi-author comment section (full list in Appendix B) independently generate comments in parallel. Each LLM receives the article, a campaign direction, and a persona with its writing-style profile drawn from the shared 400-account pool (Appendix A), but no inter-comment context or shared strategy. From up to 5 candidates per model (up to 55 across the 11 models), 20 are selected via model-balanced round-robin sampling. Engagement scores and posting timestamps are drawn from empirical Reddit distributions (Appendix B, Stages 4–5).

3.2 MACM: Coordinated Manipulation

MACM (Multi-Agent Coordinated Manipulation) adopts a vertical (plan-then-execute) architecture (Figure 1). (i) *Campaign Planning*: a single LLM call designs a 20-comment campaign specifying the narrative arc, convergence pattern, per-agent argumentation strategy, and engagement role: 3 heroes, 2–4 villains (weak opposition for heroes to refute), remaining supporters, and 1–3 background agents. Each planned role slot is then bound to a persona from the MACM 400-account pool (Appendix A). (ii) *Sequential Generation*: each agent generates one comment following the plan, receiving the last 5 preceding comments as context, the key difference from LGCM’s independent generation. We use 6 frontier models round-robin (full list in Ap-

pendix C). (iii) *Empirical Calibration*: engagement scores and posting timestamps are drawn from distributions measured across 606 Reddit posts; critically, heroes draw from upper-quartile upvote distributions, ensuring they appear at the top of the comment section, while villains draw from lower-visibility distributions; coordination thus operates on both content and visibility.

We design this architecture from the perspective of a campaign operator, deliberately selecting persuasion strategies grounded in prior persuasion research. MACM’s villain–hero debate structure creates the appearance of organic disagreement: villain agents present weak opposing arguments that heroes refute with compelling rebuttals. This design operationalizes *refutational two-sided messaging*, shown to outperform one-sided appeals (Allen, 1991; O’Keefe, 1999), and is consistent with *inoculation theory* (McGuire, 1961), where pre-exposure to weakened counterarguments strengthens the persuasive impact of subsequent rebuttals. The hero majority and upper-quartile upvote placement target social proof (Cialdini, 1984) and peripheral-route processing (Petty and Cacioppo, 1986). The 20-comment thread follows one of three convergence patterns, narrative strategies that stage apparent disagreement and determine where the hero comments appear, so that the thread converges on the campaign position while still reading as an organic discussion: *Snowball*, *Debate then Converge*, and *False Balance* (details in Appendix C).

For each article, generating all 20 comments costs approximately \$0.02 and takes 30 s with LGCM, compared with \$0.15 and 194 s with MACM.

4 Experiment

Design. Answering RQ1–RQ4 requires comparing how readers respond to different comment sources (Human vs. independently generated AI vs. coordinated multi-agent AI) and whether AI labels mitigate any persuasion effects. Because exposing one participant to multiple sources would contaminate later baselines, and seeing AI labels once would prime suspicion in subsequent unlabeled conditions, we adopted a 3 (Comment Source: Human, LGCM, MACM) $\times$ 2 (AI Label: Present, Absent) between-subjects design. Each participant was assigned to one of six conditions and read 5 news articles with 20 comments from that condition, yielding 5 article-level observations per participant.

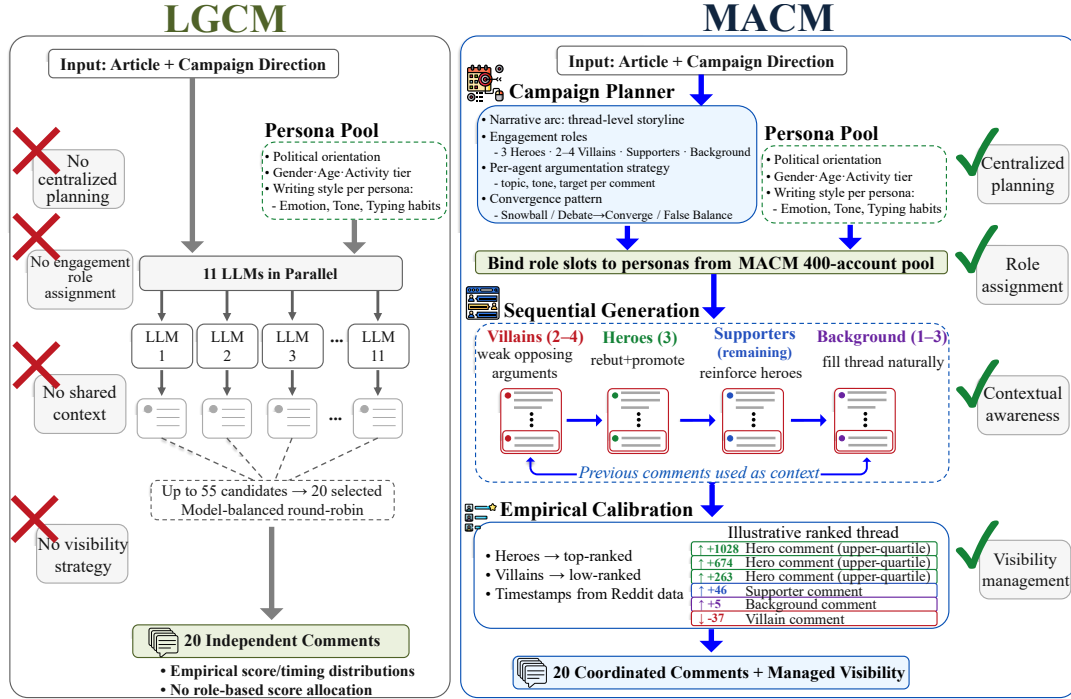


Figure 1: The two AI comment-generation pipelines. LGCM generates up to 55 candidate comments in parallel across 11 LLMs with no shared context, then selects 20 by model-balanced round-robin sampling. MACM adds the coordination package: campaign-level planning, engagement-role assignment, sequential generation in which each agent sees the last five comments, and role-aligned visibility management, yielding one coordinated 20-comment thread. Both draw on separate 400-account persona pools and calibrate engagement scores and timestamps from empirical Reddit distributions, so the LGCM-vs-MACM contrast isolates the coordination package. Full specifications appear in Appendices B and C.

Participants. Our final sample comprised 1,000 U.S. adults recruited through CloudResearch’s Connect platform (CloudResearch, 2026). Each participant was paid \$6.00 for a study averaging about 30 minutes (roughly \$12 per hour), at or above the platform’s recommended compensation guidance for U.S.-based participants. Using quota sampling (Pew Research Center, 2019), we balanced the sample at 30% Republicans, 30% Democrats, and 40% Independents, with 50/50 gender; remaining demographics (age, race, region, education, employment) were drawn via Census Match (demographics in Appendix L). The target sample size was determined a priori to detect effects as small as $d = 0.15$ (the lower bound reported in prior AI persuasion studies (Potter et al., 2024)) with 80% power ($\alpha = .05$). Each participant read 5 articles, yielding 5,000 article-level observations, with 165–168 participants per cell. Random assignment produced balanced groups on all demographic and pre-survey variables except education ($p = .047$; Table 22).

Articles. We used 80 real news articles from Reddit r/news, one of the largest English-language news discussion communities with threaded comments, upvotes, and popularity sorting, collected in early 2026 via *Top posts of the past month* (collection dates in Appendix E). We retained articles with at least 50 top-level comments, then two researchers selected 80 balanced across Economy/Society/Controversy/Politics. Outlet bias was controlled by excluding outlets rated Left or Right by AllSides Media Bias Ratings (AllSides, 2025). Each participant was randomly assigned five articles, with all four categories represented (Appendix E).

Study environment. Participants interacted with a custom web application that replicates Reddit’s interface (threaded comments, upvotes/downvotes, sort controls, reply functionality) to ensure ecological validity without field deployment. Comments appeared in popularity-sorted (“top”) order by default, matching real Reddit norms. Screenshots of the study environment are provided in Appendix K.

Comments. Because each participant is assigned to exactly one of six conditions, all 20 comments accompanying each article come from a single source (Human, LGCM, or MACM). Human comments are the first 20 top-level comments by posting time from each article’s original Reddit thread, preserving authentic discourse. AI comments are pre-generated for 36 condition variants per article, combining 7 baseline opinion levels (each participant’s pre-comment agreement rating, 1–7 Likert; see §4), 3 political orientations (progressive/moderate/conservative), and 2 campaign directions (promote or suppress the article’s claim), minus 6 infeasible combinations. Each participant was *dynamically routed* to the variant matching their pre-comment baseline and political orientation. This mirrors the audience segmentation real campaign operators perform. In Label conditions, flagged comments were displayed with the AI-label badge described below. Comment-level corpus statistics (length, lexical diversity, readability, and stylistic markers) are reported in Appendix F.

AI labeling. In the Label condition, each comment classified as AI-generated is marked with a visible “Likely AI-generated” badge so that readers can see the classification while reading comments. Classification is performed by XDAC (Go et al., 2025), an AI-comment detector that distinguishes human-written from LLM-generated English news comments, built as a GPT-4o classifier fine-tuned on 400K English comments ($F1 = 97.6\%$ on the published benchmark). Evaluated on all 116,800 comments used in our experiment (1,600 human + 57,600 LGCM + 57,600 MACM), $F1 = 90.2\%$ and 2.69% of human-written comments were incorrectly flagged as AI, with LGCM/MACM recall = 83.4%/81.0% (full breakdown in Appendix D). Using a real detector with imperfect recall, rather than oracle labels, mirrors realistic deployment conditions.

Human-condition direction labeling. Human Reddit comments lack an explicit campaign direction. We classified each article’s 20 Human comments as promote/suppress/ambiguous using a six-LLM ensemble (Fleiss’ $\kappa = 0.68$ across the six models), then had two researchers, one an author and one external to the project, verify every label while coding independently of each other. The resulting distribution is 43 promote, 33 suppress, and 4 ambiguous; the 4 ambiguous articles carry no signed direction and are therefore excluded

from Human-condition analyses of directed opinion change (details in Appendix E).

Procedure. The study comprised four phases (about 30 min total). (1) *Pre-survey*: political orientation, comment trust, and news habits. (2) *Article reading* (Baseline): participants read all 5 articles, wrote their own comment on each, and rated agreement (hereafter Agr) on a 7-pt Likert (“How much do you agree with the main claim of this article?”) before seeing any comments; all baseline ratings were collected before phase 3 to prevent carryover. (3) *Comment exposure* (Treatment): the same 5 articles were re-presented in a different random order with 20 condition-assigned comments visible; participants wrote another comment, re-rated Agr, and completed a 6-item comment evaluation (including persuasiveness and believability; full item list in Appendix J.3). (4) *Post-survey*: AI perception scale, manipulation awareness (open-ended), demographics, and attention checks. To prevent priming, participants were not informed that any comments were AI-generated until the full debriefing after completion.

Measures. The primary dependent variable is **opinion change**: the average change in 7-point agreement, $\Delta Agr = s (Agr_{\text{treatment}} - Agr_{\text{baseline}})$, where $s = +1$ for promote and $s = -1$ for suppress, so that positive values indicate movement toward the campaign goal. We also report the *change rate*: the proportion of readers whose opinion moved at least one step toward the campaign direction ($\Delta Agr > 0$). Secondary measures include perceived persuasiveness ($Q_{\text{Treat.4}}$), believability ($Q_{\text{Treat.5}}$), and writing quality ($Q_{\text{Treat.6}}$), all 7-pt Likert scales administered per article after comment exposure (Appendix J.3). For RQ3, we measure opinion diversity as the cross-participant variance of treatment-phase Agr; diversity change is $\Delta Var = (Var(\text{Treatment}) - Var(\text{Baseline})) / Var(\text{Baseline})$, where negative values indicate diversity reduction.

Analysis. Because each participant contributes five article-level observations and each article is read by many participants, we model binary outcomes with Generalized Estimating Equations (GEE) and continuous outcomes with linear mixed-effects models (LMM) that include crossed random intercepts for participant and article. Two-sample comparisons use Welch’s *t*-test (Delacre et al., 2017); variance-ratio and mediation tests use participant-clustered bootstrap (Cameron and

Source	Opinion change		Opinion diversity		
	Change	d vs. Human	Var (Baseline)	Var (Treatment)	Δ Var (%)
Human	+0.02	N/A	2.58	2.47	-4.3
LGCM	+0.29	0.19	2.58	2.60	+0.7
MACM	+0.27	0.16	2.70	2.38	-11.9

Table 1: Opinion change and opinion diversity by comment source (No Label condition; $N \approx 830/\text{cell}$). *Opinion change* is the average change in 7-point agreement toward the campaign direction. Cohen’s d compares each AI source with the Human baseline. Negative variance change indicates reduced opinion diversity.

Miller, 2015) (5,000 resamples, bias-corrected accelerated 95% CI). Because promote and suppress campaigns differ structurally (§3.2), label effects are pre-specified at the Source $\times$ Direction decomposition; the four cells are tested under the same LMM with Benjamini–Hochberg false-discovery-rate correction (Benjamini and Hochberg, 1995). Opinion change is zero-inflated (47% at zero), collapsing medians to zero in all cells. We therefore report means with participant-clustered bootstrap 95% CIs (5,000 iterations) complemented by change rates. All parametric inferences are verified with matched nonparametric tests, which reproduce the parametric conclusions (Appendix I).

5 Results

5.1 RQ1: AI Comments Shift Opinions Toward the Campaign

AI comments shifted reader opinions toward the campaign direction. In the NoLabel condition, both AI sources produced mean opinion changes of +0.29 (LGCM) and +0.27 (MACM) on a 7-point scale (Cohen’s $d = 0.19$ and 0.16 vs. Human; both $p < .005$), while organic human comments produced near-zero movement (+0.02; Table 1; full 3×2 descriptives in Table 11, Appendix I). The Human condition serves as a real-world status-quo baseline rather than a counterfactual to coordinated AI; the comparison measures the marginal impact of introducing AI campaigns into existing discourse (see Limitations). This translates to 32.9% of AI-exposed readers shifting toward the campaign goal, compared with 24.9% under organic comments (+8.0 pp, $p < .001$).

AI exposure did not increase the overall *rate* of opinion change (49.9% vs. 53.0%, $p = .098$); rather, it biased the *direction*: among changers, 62.1% shifted in the campaign direction versus

49.8% under organic comments ($p < .0001$).

This aggregate is not uniformly distributed. Among readers whose baseline opinion opposed the campaign direction (OPPOSITE), 51.5% shifted toward the campaign goal; among already-aligned readers (SAME) only 18.1%, and among neutral readers (NEUTRAL) 21.9%. The OPPOSITE subgroup accounts for the bulk of the aggregate effect. For reference, organic Human comments produced OPPOSITE reversion (46.4% change rate under No-Label), establishing a natural baseline; LGCM (48.2%) was comparable, but MACM (55.2%) exceeded the Human baseline by +8.8 pp, indicating a dissenter-conversion effect unique to coordinated AI.

5.2 RQ2: Coordination Is Stronger on Dissenters

Coordination was stronger among readers who initially opposed the campaign, despite producing little difference in the whole-sample average. On the whole sample, MACM and LGCM produced nearly identical average opinion changes (+0.27 vs. +0.29; $p = .781$; Table 1). This null result is robust to nonparametric and equivalence tests (Appendix I). However, this null masks a divergence between readers who initially agreed versus disagreed.

When restricted to the OPPOSITE subgroup (readers whose baseline opinion opposes the campaign direction), MACM is significantly stronger than LGCM when pooled across Promote and Suppress directions (Cohen’s $d = +0.17$, $p = .028$). Under Promote, MACM opinion change = +1.50 (58.2% change rate) vs. LGCM +1.19 (53.3%); under Suppress, MACM +0.92 (53.4%) vs. LGCM +0.73 (45.5%). The whole-sample null arises because MACM’s larger OPPOSITE shift is offset by stronger movement among SAME readers in the opposite direction (Appendix H).

5.3 RQ3: Only Coordinated AI Reduces Opinion Diversity

Only MACM significantly reduced opinion diversity (−11.9%; participant-clustered bootstrap 95% variance-ratio CI [0.83, 0.93], excluding 1.0; Table 1). LGCM produced no significant change ([0.94, 1.07]). The reduction is robust to permutation tests (Appendix I).

A variance decomposition identifies the mechanism: MACM’s $\text{Cov}(\text{baseline}, \text{shift})$, where shift is $\text{Agr}_{\text{treatment}} - \text{Agr}_{\text{baseline}}$, was 37% more

negative than LGCM's (-1.362 vs. -0.991 ; participant-clustered bootstrap of the difference, $+0.371$, 95% CI $[0.147, 0.604]$). The decomposition is alignment-stratified: among baseline-opposed readers, *both* AI sources compressed variance (LGCM -27.0% , MACM -33.6%); among already-aligned readers, only MACM compressed (-15.7%) while LGCM expanded ($+13.1\%$). Under LGCM the two effects offset, leaving aggregate diversity intact; under MACM both subgroups compress, producing the 11.9% reduction.

5.4 RQ4: Labels Fail, a Knowing-Resisting Dissociation

AI labels raised source recognition but did not reliably reduce persuasion. Labels consistently reduced perceived believability in all four Source $\times$ Direction cells (Table 2); perceived persuasiveness also declined in direction but did not survive correction. Yet labels reduced opinion change in only one cell, left two unchanged, and increased it in a fourth. Because promote and suppress campaigns engage structurally distinct alignment mechanisms (§3.2), label effects are analyzed at the Source $\times$ Direction decomposition (§4). Across all AI observations, the overall label effect on direction-consistent change rates was null ($33.2\% \rightarrow 32.7\%$; participant-clustered GEE, OR = 0.97, 95% CI $[0.83, 1.15]$, $p = .76$). Excluding the five Label-condition participants who did not notice the labels (label-icon hover or post-survey self-report) gives the same conclusion (32.6% ; OR = 0.97, $p = .74$).

Direction-level label effects. Label effects on directed opinion change, a continuous outcome, were analyzed by Source $\times$ Direction with crossed random-intercept LMMs and Benjamini–Hochberg correction across the four cells (Table 2). Of the four cells, only LGCM $\times$ Promote shows a label-induced reduction in AI persuasion ($d = -0.19$, $p_{\text{BH}} = .030$); two cells show no label effect (LGCM $\times$ Suppress, MACM $\times$ Promote); and MACM $\times$ Suppress shows a label-induced *increase* ($d = +0.19$, $p_{\text{BH}} = .030$). Alignment-stratified decomposition (Appendix H) reveals this MACM $\times$ Suppress increase is *not* a uniform backfire: in the OPPOSITE subgroup where the suppress campaign actually succeeds, labels do **not** attenuate MACM persuasion (NoLabel $+0.92 \rightarrow$ Label $+0.97$); the pooled increase is driven by labels reducing the resistance of SAME and NEUTRAL readers, who, under NoLabel, were drifting *against* the campaign

Source	Campaign direction	Opinion change		Label effect		
		No Label	Label	Opinion change	Believability	Persuasiveness
LGCM	Promote	+0.41	+0.15	-0.19	-0.34	-0.15
	Suppress	+0.18	+0.19	+0.01	-0.30	-0.03
MACM	Promote	+0.34	+0.35	+0.01	-0.33	-0.14
	Suppress	+0.19	+0.48	+0.19	-0.26	-0.14

Table 2: Label effects by source and campaign direction ($n = 399\text{--}437/\text{cell}$). Opinion change is the average change on a 7-point agreement scale toward the campaign direction. Label-effect columns report Cohen's d comparing Label vs. No Label; positive values mean labels increased the outcome. Bold indicates $p_{\text{BH}} < .05$.

direction. In sum, labels failed to reduce opinion change in 3 of 4 cells; in MACM $\times$ Suppress, they additionally eroded the resistance of SAME and NEUTRAL readers without protecting OPPOSITE readers. Perceived believability showed consistent label-induced reductions across all four cells ($d = -0.26$ to -0.34 , all $p_{\text{BH}} < .001$); persuasiveness effects were also negative but smaller ($d = -0.03$ to -0.15) and did not survive correction, reinforcing the perception–behavior dissociation.

Convergent evidence. A signal-detection analysis of post-survey AI identification accuracy shows that under NoLabel, humans discriminate AI from human comments essentially at chance ($d' = 0.36$); the XDAC badge raises this nearly threefold ($d' = 1.07$; both AI sources $+ \sim 33$ pp, χ^2 test, $p < .001$). Yet behavior does not move: among AI-condition participants, those who reported that quite a few or most comments seemed AI-written rated comments as less believable ($d = -0.53$, $p < .001$) but were no less persuaded than those who reported none (change rate 31.8% vs. 34.0%, χ^2 test, $p = .39$). Two further measurements reinforce this dissociation: self-reported non-change (66% claimed unchanged yet actually shifted $+0.16$) and AI familiarity within AI conditions (tool experience $d = +0.29$, $p = .040$; affect $d = +0.21$, $p = .011$; the more familiar with AI, the *more* susceptible; within the high-familiarity subgroup, higher comment-reading engagement further amplifies persuasion, $d = +0.15$, $p = .009$). Source-level detectability comparisons and full five-measurement breakdown in Appendices D and H. A sixth measurement targets label comprehension directly: 69% of Label-condition participants reported that the label *decreased* their trust in labeled comments ($Q_{\text{Post.29b}}$; $M = 1.99$ on a 5-point scale where 3=no change). Yet those who reported the largest

trust reduction ($Q_{\text{Post},29} \geq 5$, $n = 233$) were not significantly less persuaded than those who reported minimal reduction (≤ 3 , $n = 61$; $d = 0.08$, $p = .57$; Appendix N).

Robustness. The LGCM $\times$ Promote mitigation vs. MACM $\times$ Suppress amplification asymmetry is robust (z -test, $p < .0001$), with a larger gap among strong-opinion readers. XDAC's human false-positive rate (2.69%) had no detectable effect on opinion or quality ratings (all $p > .1$). An article-level analysis further confirms the dissociation: across Label-condition articles, the number of comments flagged by XDAC (range 8–20; mean 16.3) does not predict opinion change (Spearman $r_s < .03$, both $p > .49$ for LGCM and MACM), confirming that exposure to more AI labels does not modulate susceptibility.

5.5 Exploratory Diagnostics

Five exploratory moderator analyses converge on the same consistent susceptibility pattern: older adults (60+), Republicans, and readers with extreme baseline opinions were more susceptible, while education, income, and neutral pre-opinions provided no buffer. Topic and article-tone effects further modulate the campaign's reach. Full statistics, article-tone results, and exploratory pre-exposure moderators appear in Appendix H.

6 Discussion

Directional persuasion in comment ecosystems. RQ1 extends AI-persuasion evidence from single-message exposure and one-on-one chatbot dialogue to a passive *comment ecosystem* with multiple voices, social-proof cues, and popularity ranking. The key implication is that AI comments can shift opinions toward a campaign direction without requiring direct AI–reader interaction. In the NoLabel AI conditions, AI comments produced average opinion changes of +0.29 (LGCM) and +0.27 (MACM) on a 7-point scale; this translated to 32.9% of readers shifting toward the campaign direction, rising to 51.5% among those who initially disagreed (§5.1). The effect was not uniform: exploratory diagnostics suggest greater susceptibility among older adults, Republicans, and readers with extreme baseline opinions, while education and income provided no detectable buffer (§5.5). The same asymmetry appears in a measure that does not rely on self-report: the comments participants wrote themselves shifted

toward the campaign direction after AI exposure but not after human exposure (Appendix G).

Dissenter conversion and diversity compression. Coordination produced no extra average shift, but it changed the structure of persuasion in two ways unique to MACM: stronger conversion of dissenters and compression of opinion diversity. One plausible explanation for the dissenter effect is that MACM assigns coordinated roles (heroes, villains, supporters) within a planned campaign, with hero comments receiving upper-quartile upvote scores that place them at the top of the thread. This coordination between content roles and social-proof ranking may be more persuasive to dissenters than LGCM's 20 independent pro-direction statements, which lack such structural reinforcement. The diversity compression follows from the same asymmetry: MACM compresses variance in both OPPOSITE and SAME subgroups, while LGCM's SAME expansion offsets its OPPOSITE compression (§5.3). Prior work has documented diversity loss in human-to-human (Lorenz et al., 2011) and LLM-to-LLM (Chuang et al., 2024) settings; to our knowledge, our 11.9% reduction provides the first empirical evidence that coordinated AI-to-human influence produces the same effect in a comment ecosystem, realizing the pathway Schroeder et al. (2026) warn of. Crucially, coordination narrows the opinion distribution without increasing the average shift (§5.2, 5.3), manufacturing apparent consensus where genuine disagreement previously existed. This manufactured uniformity undermines the diversity prerequisite for collective intelligence (Surowiecki, 2004), posing a direct threat to the quality of public deliberation.

Disclosure fails to mitigate AI persuasion. AI-transparency regulation (EU AI Act Article 50, U.S. transparency bills) rests on an implicit chain: inform users that content is AI-generated, enable informed judgment, and thereby reduce susceptibility to manipulation. Our findings break this chain at the final link: source recognition does not reliably become behavioral resistance. Although AI labels consistently reduced perceived believability, they did not reliably reduce perceived persuasiveness or *actual* opinion change; labels reduced opinion change in only one cell, left two unaffected, and increased it in the fourth (§5.4). Five convergent measures corroborate this *Knowing-Resisting dissociation*, and article-level analysis rules out insufficient label coverage as an explanation. Notably, 84.5% of Label-condition participants stated that platforms

should indicate AI-generated comments, and 81.5% of supporters preferred per-comment labeling, the very format deployed in our study (Appendix M). Users thus endorse the policy whose behavioral efficacy our data show to be limited.

The MACM $\times$ Suppress cell illustrates this dissociation most sharply: labels did not attenuate persuasion among OPPOSITE readers (NoLabel $+0.92 \rightarrow$ Label $+0.97$), while weakening the active counter-mobilization of SAME and NEUTRAL readers who, under NoLabel, were drifting against the campaign (Suppress SAME: $-1.27 \rightarrow -0.89$; NEUTRAL: $-0.20 \rightarrow -0.02$; Appendix H.3), suggesting that labels shifted these readers from active counterarguing to discounting. AI familiarity further deepened this dissociation: participants with more tool experience and positive affect toward AI were *more*, not less, susceptible (tool experience $d = +0.29$; affect $d = +0.21$; Appendix N), and high-familiarity readers who engaged more with the comments were persuaded further ($d = +0.15$, $p = .009$). Routine use of AI tools may lead readers to treat AI-generated comments as ordinary online discourse rather than as a manipulation cue, weakening the link between disclosure and resistance; familiarity promotes engagement rather than skepticism. Per-comment labels are thus insufficient against coordinated AI campaigns and carry unintended side effects on readers who would otherwise resist.

Layered mitigation beyond per-comment labels. Aggregate evaluations of AI labels can obscure meaningful variation. The gap between perception and behavior was not uniform across Source $\times$ Direction cells (§5.4). Prior null findings (e.g., Gallegos et al., 2026) may similarly reflect cell-level heterogeneity rather than uniform absence. We therefore recommend that future research adopt direction-level decomposition as primary inference, and re-examine existing null findings at the cell level. Per-comment labels can be safely deployed (no false-positive harm; §5.4) but do not reliably reduce opinion change. Our study evaluated only a single label presentation form, specifically a per-comment “Likely AI-generated” badge displayed inline; stronger forms that visually segregate AI-flagged content or convey campaign-level coordination warnings may produce different behavioral effects and remain future work. Appendix O outlines a Mitigation Framework grounded in our findings: beyond per-comment disclosure (L1, tested), platforms can deploy exposure restrictions and

social-proof stripping on flagged comments (L2), campaign-level coordination warnings when directional anomalies are detected (L3), and diversity injection that surfaces existing dissenting voices to break manufactured social proof (L4). These interventions act on the comment-section surface; complementary operator-side mechanisms (account-level coordination analysis, persistent attribution) are beyond the scope of this user study and remain future work. As a standalone mitigation, however, per-comment labeling proves ineffective or counterproductive in the very cells where coordinated campaigns matter most.

7 Conclusion

A controlled experiment with 1,000 U.S. adults yielded three findings: (i) AI comments shift opinions toward the campaign direction, especially among those who initially disagreed; (ii) multi-agent coordination persuades more of those who initially disagreed and compresses opinion diversity, manufacturing apparent consensus where genuine disagreement existed; (iii) AI labels reduce perceived believability yet fail to reduce actual opinion change, a *Knowing-Resisting dissociation*. These results provide initial evidence that coordinated AI Swarms (Schroeder et al., 2026) produce qualitatively different persuasion patterns from single-message exposure, and indicate that per-comment disclosure is necessary but insufficient on its own.

Limitations

Our experiment demonstrates that even a single exposure to 20 AI comments shifts one in three readers; however, real-world news consumption involves repeated exposure across days and platforms. Whether $d \approx 0.17$ cumulates with repetition (as in drip-feed propaganda) or attenuates through habituation is an important open question. We hope our findings motivate larger-scale field experiments that track longitudinal exposure effects.

The Label condition deliberately used imperfect detection (XDAC recall: 83% for LGCM, 81% for MACM; Appendix D) to mirror deployable platform conditions rather than an idealized oracle. A supplementary oracle-label condition would have enabled disentangling detection gaps from label effectiveness in future work.

The organic Human condition reflects mixed-stance Reddit threads, the real-world status quo that readers currently encounter, while AI conditions model directed campaigns. The Human-vs-AI comparison therefore measures the marginal impact of introducing coordinated AI into existing discourse, not whether AI is inherently more persuasive than equivalent human-authored directed campaigns. Our central coordination findings (RQ2–RQ3) are further isolated from this distinction by the LGCM-vs-MACM contrast, which holds generator type and direction constant. A human-authored directed-campaign condition would address a separate question about generator-type effects within directed campaigns, and we leave this to future work.

Similarly, we did not evaluate a cost-matched baseline (e.g., LGCM with $7\times$ more comments), as the increased volume introduces confounds from visibility and reader fatigue. Our MACM condition intentionally evaluates coordination as a system-level package: centralized planning, role assignment, contextual awareness, and visibility management operate together. The stance composition and engagement-score differences between LGCM and MACM are therefore realized effects of this package rather than separate confounds. Isolating each component would require a substantially larger factorial design. Our article corpus spans approximately two months, from early January through late February 2026, which may limit the temporal generalizability of our findings to other news cycles. Finally, all participants were English-speaking U.S. adults; generalization to other languages, cul-

tures, and platform ecosystems (e.g., Weibo, Naver) requires replication.

Ethical Considerations

We recognize that the MACM pipeline could be misused for real-world opinion manipulation. However, the threat of coordinated AI-generated comments already exists via commercially available tools, and empirical evidence of its effectiveness is necessary for policymakers and platform designers to develop countermeasures; withholding evidence would leave defenders uninformed while adversaries proceed regardless. Accordingly, we document the stimulus-generation procedure and detector evaluation needed to reproduce and audit our findings, but withhold full prompt templates, per-account orchestration details, and civil-account acquisition tooling that would lower the operational barrier to real-world deployment. This study was approved by our institution's Institutional Review Board. All participants gave informed consent before starting: the consent form stated the approximate duration and compensation, that participation was voluntary and could be discontinued at any time without penalty, and that their responses would be used for research and reported only in aggregate. Because the experiment involved deception, a thorough debriefing was conducted after task completion; participants could withdraw their data, and none did.

The Human-condition stimuli are real comments from public Reddit threads. Before using them we discarded all author usernames and account metadata and masked personally identifying information, including personal names, contact details, account handles, and links to identifiable accounts, so that participants saw only this anonymized form. We deliberately did not filter offensive language beyond this masking, because the Human condition is meant to reproduce the unmoderated discourse that readers actually encounter, and removing it would have made the baseline unrepresentative.

Acknowledgments

Part of this work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (Nos. RS-2024-00436936 and RS-2022-II220688), awarded to Hyoungshick Kim (third author).

References

- Mike Allen. 1991. [Meta-analysis comparing the persuasiveness of one-sided and two-sided messages](#). *Western Journal of Speech Communication*, 55(4):390–404.
- AllSides. 2025. [AllSides Media Bias Ratings](#). Version 11, released 2025-12-08; accessed 2026-01.
- Anthropic. 2025. [Introducing Claude 4](#).
- Solomon E. Asch. 1956. [Studies of independence and conformity: I. A minority of one against a unanimous majority](#). *Psychological Monographs: General and Applied*, 70(9):1–70.
- Hui Bai, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. 2025. [LLM-generated messages can persuade humans on policy issues](#). *Nature Communications*, 16:6037.
- Yoav Benjamini and Yosef Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.
- A. Colin Cameron and Douglas L. Miller. 2015. [A practitioner’s guide to cluster-robust inference](#). *Journal of Human Resources*, 50(2):317–372.
- Damon Centola and Michael Macy. 2007. [Complex contagions and the weakness of long ties](#). *American Journal of Sociology*, 113(3):702–734.
- Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. 2024. [Simulating opinion dynamics with networks of LLM-based agents](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346.
- Robert B. Cialdini. 1984. *Influence: How and Why People Agree to Things*. William Morrow, New York.
- CloudResearch. 2026. [Connect: Premier participant recruitment platform](#). Accessed: 2026-08-27.
- Michael A. Covington and Joe D. McFall. 2010. [Cutting the Gordian knot: The moving-average type–token ratio \(MATTR\)](#). *Journal of Quantitative Linguistics*, 17(2):94–100.
- Marie Delacre, Daniël Lakens, and Christophe Leys. 2017. [Why psychologists should by default use Welch’s t-test instead of Student’s t-test](#). *International Review of Social Psychology*, 30(1):92–101.
- Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. [Measuring the persuasiveness of language models](#). Anthropic Research Post.
- Bradley Efron and Robert J. Tibshirani. 1993. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.
- Isabel O. Gallegos, Chen Shani, Weiyan Shi, Federico Bianchi, Izzy Gainsburg, Dan Jurafsky, and Robb Willer. 2026. [Labeling messages as AI-generated does not reduce their persuasive effects](#). *PNAS Nexus*, 5(2):pgag008.
- Wooyoung Go, Hyoungshick Kim, Alice Oh, and Yongdae Kim. 2025. [XDAC: XAI-driven detection and attribution of LLM-generated news comments in Korean](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22728–22750, Vienna, Austria. Association for Computational Linguistics.
- Josh A. Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. [How persuasive is AI-generated propaganda?](#) *PNAS Nexus*, 3(2):pgae034.
- Google DeepMind. 2025. [Gemini 2.5: Our most intelligent AI model](#).
- Shiyang Lai, Junsol Kim, Nadav Kunievisky, Yujin Potter, and James Evans. 2025. [Biased AI improves human decision-making but reduces trust](#). *arXiv preprint arXiv:2508.09297*. NeurIPS 2025 Workshop on Algorithmic Collective Action.
- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, 33(1):159–174.
- Hause Lin, Gabriela Czarnek, Benjamin Lewis, Joshua P. White, Adam J. Berinsky, Thomas Costello, Gordon Pennycook, and David G. Rand. 2025. [Persuading voters using human–artificial intelligence dialogues](#). *Nature*, 648(8093):394–401.
- Teng Lin and Yiqing Zhang. 2026. [Visible sources and invisible risks: Exploring the impact of AI disclosure on perceived credibility of AI-generated content](#). *Journal of Science Communication*, 25(1):A09.
- Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. [How social influence can undermine the wisdom of crowd effect](#). *Proceedings of the National Academy of Sciences*, 108(22):9020–9025.
- Weicheng Ma, Hefan Zhang, Ivory Yang, Shiyu Ji, Joice Chen, Farnoosh Hashemi, Shubham Mohole, Ethan Gearey, Michael Macy, Saeed Hassanpour, and Soroush Vosoughi. 2025. [Communication makes perfect: Persuasion dataset construction via multi-LLM communication](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4017–4045, Albuquerque, New Mexico. Association for Computational Linguistics.
- William J. McGuire. 1961. [The effectiveness of supportive and refutational defenses in immunizing and restoring beliefs against persuasion](#). *Sociometry*, 24(2):184–197.

- Elisabeth Noelle-Neumann. 1974. [The spiral of silence: A theory of public opinion](#). *Journal of Communication*, 24(2):43–51.
- Daniel J. O’Keefe. 1999. [How to handle opposing arguments in persuasive messages: A meta-analytic review of the effects of one-sided and two-sided messages](#). *Annals of the International Communication Association*, 22(1):209–249.
- OpenAI. 2024. [GPT-4o system card](#).
- OpenAI. 2025. [Introducing GPT-5](#).
- Richard E. Petty and John T. Cacioppo. 1986. [The elaboration likelihood model of persuasion](#). In Leonard Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 19, pages 123–205. Academic Press.
- Pew Research Center. 2019. [Political independents: Who they are, what they think](#). Accessed: 2026-04-30.
- Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. [Hidden persuaders: LLMs’ political leaning and their influence on voters](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4244–4275, Miami, Florida, USA. Association for Computational Linguistics.
- r/changemyview moderators. 2025. [META: Unauthorized experiment on CMV involving AI-generated comments](#). University of Zurich field study disclosed 2025-04-26; the authors subsequently withdrew the planned publication.
- Francesco Salvi, Manoel Horta Ribeiro, Riccardo Galotti, and Robert West. 2025. [On the conversational persuasiveness of GPT-4](#). *Nature Human Behaviour*, 9(8):1645–1653.
- David Schoch, Franziska B. Keller, Sebastian Stier, and JungHwan Yang. 2022. [Coordination patterns reveal online political astroturfing across the world](#). *Scientific Reports*, 12:4572.
- Daniel Thilo Schroeder, Meeyoung Cha, Andrea Baronchelli, Nick Bostrom, Nicholas A. Christakis, David Garcia, Amit Goldenberg, Yara Kyrychenko, Kevin Leyton-Brown, Nina Lutz, Gary Marcus, Filippo Menczer, Gordon Pennycook, David G. Rand, Maria Ressa, Frank Schweitzer, Dawn Song, Christopher Summerfield, Audrey Tang, Jay J. Van Bavel, Sander van der Linden, and Jonas R. Kunst. 2026. [How malicious AI swarms can threaten democracy](#). *Science*, 391(6783):354–357.
- Skipper Seabold and Josef Perktold. 2010. [Statsmodels: Econometric and statistical modeling with Python](#). In *Proceedings of the 9th Python in Science Conference*, pages 92–96.
- Tianqi Song, Yugin Tan, Zicheng Zhu, Yibin Feng, and Yi-Chieh Lee. 2025. [Multi-agents are social groups: Investigating social influence of multiple agents in human-agent interactions](#). *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–33. Article CSCW452.
- S. Shyam Sundar. 2008. The MAIN model: A heuristic approach to understanding technology effects on credibility. In Miriam J. Metzger and Andrew J. Flanagin, editors, *Digital media, youth, and credibility*, pages 73–100. MIT Press, Cambridge, MA.
- James Surowiecki. 2004. *The wisdom of crowds*. Doubleday.
- Thomas Zerback and Florian Töpl. 2022. [Forged examples as disinformation: The biasing effects of political astroturfing comments on public opinion perceptions and how to prevent them](#). *Political Psychology*, 43(3):399–418.

A Shared Infrastructure: Persona Pool and Writing-Style Calibration

This appendix describes the persona pool and writing-style system shared by both the LGCM and MACM pipelines (Appendices B and C).

A.1 Persona pool construction

We pre-built 800 synthetic Reddit accounts, 400 for LGCM and 400 for MACM. Each account carries a fixed identity (username, karma, bio, writing style) that persists across all 80 articles, ensuring that the same persona produces stylistically consistent comments throughout the experiment. The two pools were generated independently with zero username overlap.

Pre-allocated attributes. Each account was created with four attributes balanced across the pool, so that no political orientation, gender, or age group is over-represented, thereby preventing the AI comment ecosystem from exhibiting demographic skew that would differ from a real Reddit comment section:

- **Political orientation** (5 levels, 80 accounts each): strongly conservative / conservative / moderate / liberal / strongly liberal.
- **Gender** (2 levels, 200 accounts each): male / female.
- **Age bracket** (6 levels, approximately 67 accounts each): teens / 20s / 30s / 40s / 50s / 60+.
- **Activity tier** (4 levels): newbie 30% / regular 40% / veteran 20% / power 10%.

LLM-generated attributes. Conditioned on the pre-allocated attributes, an LLM generated the remaining profile fields: username (stylistically modeled after 100 real Reddit usernames to avoid obviously artificial handles), occupation, location, bio, flair, post karma, comment karma, and account creation date. All values were constrained to fall within the ranges typical of the assigned activity tier.

Cross-pool deduplication. The LGCM pool was generated first; the MACM pool was then generated with a cross-check against the LGCM pool to ensure zero username overlap. Within each pool, duplicate usernames were regenerated until all 400 were unique.

A.2 Writing-style calibration

Reddit comment analysis. To ground AI-generated comments in realistic stylistic patterns, we analyzed 1,600 real Reddit news comments (the same 80 articles $\times$ 20 first-by-time top-level comments used as the Human condition). Using GPT-5.4 for inductive tagging, we derived a writing-style taxonomy with three trait dimensions:

- **Emotion** (9 primary categories, covering 97.8% of comments): skeptical (17.5%), sarcastic (16.8%), neutral (16.0%), cynical (11.3%), profanity (9.5%), humorous (8.8%), dismissive (7.3%), passionate (7.2%), emotional expression (3.4%).
- **Tone** (8 primary categories, covering 99.5%): informal (35.3%), rhetorical questions (17.9%), brief (16.1%), matter-of-fact (14.1%), analytical (6.8%), speculative (4.6%), personal anecdote (3.7%), formal (0.9%).
- **Typing habits** (14 categories): paragraph breaks (18.1%), quoted text (8.7%), lowercase (7.3%), ellipsis (6.3%), extra spaces (5.3%), slang (5.2%), scare quotes (3.7%), spelling errors (2.7%), all-caps (2.5%), no punctuation (1.6%), emoji (1.6%), run-on sentences (1.6%), line breaks (1.5%), parenthetical asides (1.3%).

Per-persona style assignment. Each of the 800 personas was assigned a writing-style profile via a two-stage process. LLM assignment alone tends to over-represent salient categories (e.g., “sarcastic” for young users); the code-based redistribution corrects these biases to match the empirical Reddit distribution.

1. **LLM-based initial assignment:** GPT-5.4 selected two emotion candidates, two tone candidates, and all applicable typing habits for each persona, conditioned on age, gender, occupation, bio, and political orientation (e.g., a young progressive student receives slang + emoji + informal).
2. **Code-based redistribution:** the pool-level distribution was compared against the Reddit target distribution; over-represented labels were swapped with under-represented ones until all categories converged within $\pm 3\%$ of the target.

At comment-generation time, one emotion and one tone are randomly sampled from each persona’s two candidates. Two candidates (rather than one fixed label) produce natural within-person variation because real users do not write every comment in exactly the same tone, while keeping both options persona-consistent.

B LGCM Pipeline

This appendix specifies the LGCM (LLM-Generated Comment Manipulation) pipeline introduced in §3.1. The pipeline has five stages: (1) article ingestion, (2) persona assignment, (3) independent parallel generation, (4) engagement-score simulation, (5) timestamp simulation. The persona pool and writing-style system are shared with MACM and described in Appendix A. A side-by-side comparison with the MACM pipeline is shown in Figure 1.

B.1 Stage 1: Article Ingestion

Each article is loaded with its original title, body text, and topic category (Economy / Society / Controversy / Politics), along with a manipulation direction specifying whether to *promote* or *suppress* the article’s claim.

B.2 Stage 2: Persona Assignment

For each article, 20 personas are drawn from the 400-account LGCM pool (Appendix A.1), filtered by political orientation compatible with the manipulation direction, so that a promote-direction campaign draws personas who would plausibly support the article’s claim. Each selected persona carries the pre-assigned writing-style profile (Appendix A.2).

B.3 Stage 3: Independent Parallel Generation

Using 11 LLMs across three provider families ensures stylistic diversity: a thread generated by a sin-

gle model would exhibit uniform phrasing that readers or detectors could recognize. The 11 models are OpenAI (GPT-5.1, GPT-5.2, GPT-5.4), Anthropic (Claude Opus 4.6, Claude Sonnet 4.6, Claude Opus 4.5, Claude Sonnet 4.5), and Google (Gemini 3 Flash Preview, Gemini 2.5 Pro, Gemini 2.5 Flash, Gemini 2.0 Flash).

Each LLM receives the article, the manipulation direction, the assigned persona (including writing-style instructions), and an empirically calibrated word-count target sampled from the Reddit comment-length distribution (p25–p75: 9–28 words). Each model generates up to 5 comments per batch call, producing a pool of up to 55 candidates. From this pool, 20 are selected via model-balanced round-robin sampling to guarantee representation from all families.

Independence. Each LLM generates its comment independently: it cannot see what other LLMs wrote, follows no shared campaign plan, and plays no assigned role.

B.4 Stage 4: Engagement-Score Simulation

Upvote and downvote counts are sampled from empirical distributions measured across the 606 Reddit-post calibration corpus to make the thread appear organically ranked. Unlike MACM (Appendix C.5), engagement scores are assigned by time-based position only, producing a flat upvote distribution because LGCM has no hero/villain roles to prioritize.

B.5 Stage 5: Timestamp Simulation

Posting timestamps are drawn from the same corpus’s empirical inter-arrival distribution so that comments do not appear at suspiciously regular intervals.

B.6 Cost and latency

Generating a full set of 20 comments for one article costs approximately \$0.02 in total API fees and completes in about 30 s under parallel calls across the 11 models.

C MACM Pipeline

This appendix specifies the MACM (Multi-Agent Coordinated Manipulation) pipeline introduced in §3.2. The persona pool and writing-style system are shared with LGCM and described in Appendix A. The pipeline has six stages: (1) article ingestion, (2) Campaign Planning, (3) persona instantiation,

(4) Sequential Generation, (5) engagement-score simulation, (6) timestamp simulation.

C.1 Stage 1: Article Ingestion

Identical to LGCM Stage 1 (Appendix B.1): article text, topic category, and manipulation direction are loaded as inputs to the Planner.

C.2 Stage 2: Campaign Planning

A single Planner agent (one LLM call, round-robin from the 6 frontier models in §3.2) designs a structured campaign plan comprising four elements:

- **Narrative arc:** how the 20-comment thread should evolve, including where the core argument is introduced, where opposition appears, and how the thread converges toward the operator’s position.
- **Convergence pattern:** one of three strategies: *Snowball*, *Debate then Converge*, or *False Balance* (detailed below).
- **Per-agent specification** ($\times 20$): political orientation, emotional valence, argumentation strategy (1 of 11 types), engagement role ({hero, villain, supporter, background}), and anchor claim.
- **Role distribution:** 3 heroes, 2–4 villains, supporters, and background agents, summing to 20.

Convergence Patterns. A comment thread in which all 20 agents uniformly push one direction would be easily recognized as manipulation. To simulate the organic disagreement found in real discussions, the Planner selects one of three narrative strategies, each inspired by documented influence-operation tactics (Zerback and Töpfl, 2022; Schoch et al., 2022):

- **Snowball:** 1–2 strong advocates push the narrative early, then progressively more agents join. Heroes are placed in early positions (1–10) to maximize bandwagon effect.
- **Debate then Converge:** genuine-seeming debate with opposing views, then counterarguments are systematically defeated, leading to convergence. Heroes are placed late (10–20) as decisive conclusers.
- **False Balance:** the appearance of balanced discourse throughout, but covertly favoring one side via quantity, emotional weight, and framing. Heroes are distributed evenly to avoid clustering.

Villain–Hero debate structure. A thread of 20 uniformly agreeing comments would appear suspicious to readers. The villain–hero structure addresses this by creating apparent disagreement: villain agents present weak opposing arguments that hero agents then refute with compelling rebuttals. This makes the thread read like a genuine debate that organically converges, rather than a one-sided campaign.

Prompt disclosure. The Stage-2 planning prompt is assembled from eight blocks: (A) the target article, (B) the persuasion goal and campaign direction, (C) the target user profile, (D) the politics distribution, (E) the emotion distribution, (F) the convergence pattern, (G) the engagement-role assignment, and (H) the strategy inventory. The prompt was instantiated once per (direction, baseline agreement, political orientation) cell to pre-generate the comment variants; the study application then routed each participant to the variant matching their own baseline agreement and political orientation. Blocks (C) and (D) are the two that condition on these routing variables, and we reproduce only those two, verbatim, in Figure 2. We withhold blocks (E)–(G), which specify the calibrated emotion ratios, the stage-by-stage narrative guide, and the role counts and hero-placement rules that make a campaign appear organically ranked; releasing them would supply a directly operational template, as discussed in the Ethical Considerations.

C.3 Stage 3: Persona Instantiation

Each plan entry is bound to a persona from the 400-account MACM pool (Appendix A), matched by political orientation and tone compatibility. Binding to pre-built personas rather than generating ad-hoc identities ensures that each agent writes in a consistent voice across articles. Word-count targets are sampled from separate empirical distributions for heroes (p25–p75: 11–40 words) and non-heroes (9–28 words), so that hero comments appear longer and more substantive, mimicking the pattern observed in top-voted Reddit replies.

C.4 Stage 4: Sequential Generation

For the thread to read as a coherent conversation rather than 20 isolated opinions, each agent must be aware of what has already been said. Comments are therefore produced sequentially: each agent receives the article, the campaign plan, its own persona, and the last 5 preceding comments, allowing

it to reference earlier agents by name and respond to their arguments. Models are assigned round-robin from 6 frontier LLMs (GPT-5.4, GPT-5.2, Claude Sonnet 4.5, Claude Sonnet 4, Claude Opus 4.5, Claude Haiku 4.5). LGCM uses 11 models, but 5 of them (GPT-5.1, Gemini family, Claude Opus/Sonnet 4.6) were excluded from MACM because they could not reliably maintain persona, role, argumentation strategy, and narrative consistency simultaneously across 20 sequential turns in pilot testing.

C.5 Stage 5: Engagement-Score Simulation

Real Reddit threads show a clear pattern: early, high-quality comments accumulate the most upvotes, while later or weaker comments receive fewer. To reproduce this, upvote and downvote counts are sampled from empirical distributions measured across the 606 Reddit-post calibration corpus, conditional on the agent’s engagement role: heroes draw from upper-quartile distributions to appear as top-voted comments, villains from lower-quartile distributions, and others from the median range.

C.6 Stage 6: Timestamp Simulation

Comments that arrive at perfectly regular intervals, cluster in an unnaturally short burst, or appear faster than a human could type would all be detectable signs of automation. Posting timestamps are therefore sampled from the calibration corpus’s empirical inter-arrival distribution to match realistic Reddit posting cadence.

C.7 Cost and latency

Generating a full set of 20 comments for one article costs approximately \$0.15 in total API fees and completes in about 194 s (sequential generation dominates, as each agent must wait for the preceding comment).

D XDAC Detection: In-Domain Audit

This appendix provides a detailed breakdown of XDAC’s detection performance on our experiment stimulus pool, expanding the aggregate metrics reported in §4.

D.1 Stimulus pool composition

The full stimulus pool consists of 116,800 comments: 1,600 Human (80 articles $\times$ 20 first-by-time top-level comments from the original Reddit

```

## Target User Profile
Design personas to be maximally persuasive to THIS specific user.
- Political Orientation: {user_orientation}
- Current Agreement: {user_agreement}

Strategy guidance:
- Match some agents' political orientation to the user's to build rapport
- If promoting: reinforce existing tendencies, amplify agreement
- If suppressing: introduce doubt, present compelling counter-framing
- Use the user's political orientation to calibrate which framing resonates

## Politics Distribution
Base politics distribution on user's political orientation ({user_orientation}):
- Assign majority of agents to the user's political orientation or adjacent
- This maximizes perceived credibility and in-group persuasion effect
- Include some agents from other orientations for authenticity

```

Figure 2: Blocks (C) and (D) of the Stage-2 planning prompt, reproduced verbatim. These are the two blocks that condition on the routing variables: {user_orientation} and {user_agreement} are filled with the political orientation and pre-comment baseline agreement that define each variant. LGCM and MACM share the same variant grid and routing procedure, so this targeting does not account for the MACM-versus-LGCM difference.

threads), 57,600 LGCM (80 articles $\times$ 36 condition variants $\times$ 20 comments), and 57,600 MACM (same structure). All comments were evaluated by XDAC prior to the experiment; flagged comments in the Label condition displayed the “Likely AI-generated” badge (Figure 4).

D.2 Aggregate detection metrics

Table 3 summarizes XDAC’s in-domain performance. The in-domain F1 (90.2%) is lower than the published benchmark (97.6%), reflecting both the domain shift from the original training corpus to Reddit news comments and the generation gap: XDAC was trained on comments from earlier-generation LLMs, whereas our stimuli were produced by substantially more capable frontier models that generate more human-like text. Precision remains near-perfect (99.95%), meaning virtually all comments flagged as AI were indeed AI-generated. The primary source of error is recall: approximately 1 in 5 AI-generated comments evaded detection.

Metric	Value
Published benchmark F1	97.6%
In-domain F1	90.2%
In-domain Precision	99.95%
In-domain Recall (overall)	82.2%
Human false-positive rate	2.7%
LGCM recall	83.4%
MACM recall	81.0%

Table 3: XDAC detection performance on the 116,800-comment stimulus pool.

D.3 Recall by source

MACM comments were slightly harder to detect than LGCM comments (81.0% vs. 83.4%), consistent with MACM’s greater stylistic diversity arising from persona and role variation. However, this 2.4 percentage-point gap is small, and both sources are detected at comparable rates.

D.4 Recall by MACM engagement role

To test whether certain roles (e.g., heroes with more substantive, longer comments) are systematically harder to detect, we stratified MACM recall by engagement role:

Role	Recall (%)
Hero	81.5
Villain	82.5
Supporter	80.8
Background	78.7

Table 4: XDAC recall by MACM engagement role.

Recall is uniformly distributed across roles (78–82%).

D.5 Human false positives

Of 1,600 human comments, 43 (2.69%) were incorrectly flagged as AI-generated. As reported in §5.4, this mislabeling had no detectable effect on readers’ opinion, believability, persuasiveness, or quality ratings for human comments (all $p > .1$), confirming that per-comment AI labeling can be deployed without distorting readers’ evaluation of genuine human discourse.

D.6 Recall by generating model

Because different LLMs produce stylistically distinct outputs, detection difficulty varies substantially across models. Table 5 reports LGCM recall by the 11 generating models; Table 6 reports MACM recall by the 6 generating models.

Generating model	Recall (%)
GPT-5.2	93.1
Claude Opus 4.5	91.1
Gemini 2.0 Flash	90.4
GPT-5.4	89.6
Gemini 2.5 Flash	89.4
Gemini 2.5 Pro	85.7
Gemini 3 Flash	84.6
Claude Sonnet 4.6	84.3
GPT-5.1	73.1
Claude Opus 4.6	73.0
Claude Sonnet 4.5	68.1
Overall	83.4

Table 5: XDAC recall on LGCM comments by generating model (sorted by recall, $N = 57,600$).

Generating model	Recall (%)
Claude Sonnet 4.5	90.2
Claude Opus 4.5	89.3
Claude Sonnet 4	88.9
Claude Haiku 4.5	88.4
GPT-5.2	69.1
GPT-5.4	62.7
Overall	81.0

Table 6: XDAC recall on MACM comments by generating model (sorted by recall, $N = 57,600$).

In LGCM, detection difficulty spans a 25-pp range across models, with the newest generations (Claude Sonnet 4.5: 68%, Claude Opus 4.6: 73%) being hardest to detect. In MACM, a model-family gap emerges: all four Claude models achieve 88–90% recall, while both GPT models fall to 63–69%.

E Article Corpus and Direction Labelling

E.1 Corpus construction

Source and collection. News articles were collected from Reddit’s *r/news* subreddit, one of the largest English-language news discussion communities. Two collection sweeps were conducted on 2026-02-06 and 2026-02-26 using the “*Top posts of the past month*” sort, effectively covering news from early January through late February 2026.

Posts with fewer than 50 top-level comments were excluded because the Human condition requires 20 first-by-time comments per article, and a minimum of 50 ensures adequate depth and diversity in the available pool.

Topic diversity. Four topic categories were pre-defined as Economy, Politics, Society, and Controversy, with category-specific keyword sets for searching within *r/news*. The initial search yielded 606 category-keyword matches (Economy: 104, Politics: 252, Society: 167, Controversy: 83).

Media bias filtering. To control for outlet-level political bias, we applied the AllSides Media Bias Ratings (AllSides, 2025). Only articles from outlets rated *Center* (Reuters, BBC, WSJ, The Hill, Newsweek, NewsNation) or *Lean Left* (CNN, NBC News, NPR, NYT, ABC News, CBS News, USA Today, CNBC, PBS) were retained; outlets rated *Left* or *Right* were excluded, removing 345 posts. No *Lean Right* outlets appeared among the top posts during the study period, which we attribute to the user-composition characteristics of the *r/news* community and discuss as a limitation.

Deduplication and reclassification. After bias filtering, deduplication by post ID yielded 166 unique posts (individual posts often matched keywords from multiple categories). Each post was reassigned to exactly one category based on the primary framing of the headline and article content. Economy posts centered on economic impact (employment, trade, corporate activity); Politics posts centered on governmental power (executive/legislative/judicial decisions); Society posts covered events affecting civic life (public safety, civil rights, public health, crime); Controversy was reserved for polarization issues where reader opinion divides sharply.

Final selection. Two researchers independently selected 80 articles (20 per category) to maximize topical diversity within each category: articles covering identical events were grouped, and only the most representative article was retained from each group. Per-subtopic caps (e.g., $\leq 25\%$ for ICE-related articles in Society) prevented overrepresentation of any single issue.

Final composition. The 80 selected articles span diverse subtopics within each category:

- **Economy** (20): economic indicators, corporate restructuring, trade/tariffs, energy, antitrust.

- **Politics** (20): judicial–executive conflict, immigration policy, foreign affairs, military, DOGE/fiscal.
- **Society** (20): ICE/civil rights, criminal justice, public health/welfare, child safety, AI/digital.
- **Controversy** (20): vaccine debate, gun control, DEI/gender, ICE policy, AI ethics, environmental regulation.

E.2 Human-condition direction labelling

AI comments have an explicit campaign direction (promote or suppress), but Human Reddit comments do not. To assign a direction sign for the Human condition, the overall direction of each article’s 20 Human comments was classified as *promote* / *suppress* / *ambiguous* by an ensemble of six LLMs spanning three providers: Anthropic Claude Opus 4 and Sonnet 4 (Anthropic, 2025); OpenAI GPT-4o (OpenAI, 2024) and GPT-5 (OpenAI, 2025); Google Gemini 2.5 Pro and Flash (Google DeepMind, 2025). Every ensemble label was then verified by two researchers, one an author of this paper and one a researcher external to the project, who coded independently of each other and were blind to the other’s codes; disagreements between the two verifiers were resolved by discussion.

Inter-rater agreement. The agreement statistics below characterise the *six-model ensemble*, not the two human verifiers. Fleiss’ $\kappa = 0.68$ across the six models (substantial agreement per the Landis and Koch (1977) benchmark); mean pairwise Cohen’s $\kappa = 0.677$ (range: 0.475–0.975), with no single model acting as a systematic outlier. A $\geq 5/6$ strong-majority threshold was reached on 84.5% of articles; the remaining 15.5% were resolved by simple majority with researcher arbitration on ties. Final article-direction distribution: 43 promote / 33 suppress / 4 ambiguous.

Treatment of ambiguous articles. The direction sign enters our analyses only through the direction-signed outcome ΔAgr (§4), which is undefined when no campaign direction can be assigned. The 4 ambiguous articles are therefore excluded from all Human-condition analyses of directed opinion change rather than being coded as zero; they are retained in undirected analyses that use raw agreement change. This affects the Human condition only, since LGCM and MACM comments carry an explicit assigned direction.

E.3 Full list of stimulus articles

Table 7 lists all 80 stimulus articles, 20 per topic category, numbered within category and ordered by Reddit post date. The *Outlet* column gives the publisher linked by the Reddit post; the nine outlets represented are Reuters (19 articles), NBC News (16), ABC News (11), BBC (8), NYT (8), CNN (7), NPR (6), CNBC (4), and USA Today (1). The *Reddit post date* column serves as each article’s timestamp: among the 48 articles whose source URL encodes a publication date, the post date follows publication by at most two days (median zero). Post dates run from January 6 to February 25, 2026, consistent with the two collection sweeps of February 6 and February 26 described in Appendix E.1.

F Comment Corpus Statistics

Table 8 reports comment-level corpus statistics for the three comment sources used in the study: Human, LGCM, and MACM. Promote/Suppress sub-corpora were audited separately and showed no substantively meaningful length differences (all words/comment comparisons $|d| < 0.10$), so they are collapsed here for readability.

Length. Human comments have a higher *mean* length than either AI condition (25.1 vs. 17.3 / 17.7 words), but the gap is driven entirely by a long-tail subset of very long Human comments (SD=33.5 vs. ≈ 10 for AI). The *median* comment length is essentially identical across the three pools (Human 16, LGCM 15, MACM 16), and a Mann–Whitney rank-sum test finds no significant difference between Human and either AI pool (Human–LGCM $p=.93$; Human–MACM $p=.37$). LGCM and MACM are themselves closely matched in length ($d=-0.06$), so the two AI corpora are well matched on surface length.

Lexical diversity. We omit raw total-word and raw vocabulary counts because they are mechanically dominated by pool size (Human $N=1,600$ vs. $N=57,600$ for each AI condition). Instead, after down-sampling each pool to $N=1,600$ comments and averaging across 50 bootstrap draws, Human shows the largest type inventory (6,383), with MACM (5,144) above LGCM (4,091). Corpus-level MATTR (window=100; Covington and McFall (2010)), a complementary *local* diversity measure, ranks the conditions differently: MACM produces the highest within-window type/token ratio (0.848), with Human (0.774) at the bottom. The

#	Outlet	Reddit post date	Headline
Economy			
1	Reuters	2026-01-14	US posts record \$145 billion December deficit as outlays outpace receipts
2	NPR	2026-01-14	Offshore wind developer prevails in U.S. court as Trump calls wind farms 'losers'
3	ABC News	2026-01-15	1.4 million fewer people enrolled in ACA plans as premiums spike, tax credits expire
4	CNN	2026-01-22	President Trump sues JPMorgan Chase and CEO Jamie Dimon, alleging the bank improperly dropped him as a client for political reasons
5	Reuters	2026-01-23	FDIC clears way for Ford, General Motors to set up industrial banks
6	CNBC	2026-01-26	Power prices surge as winter storm spikes demand in U.S. data center alley
7	CNN	2026-01-27	Amazon is closing its Fresh and Go locations
8	NYT	2026-01-28	Amazon to Cut 16,000 Jobs in Latest Round of Layoffs
9	NBC News	2026-01-29	Elon Musk says Tesla will stop producing its S and X models as it shifts to making robots
10	Reuters	2026-01-29	US trade deficit widens by the most in nearly 34 years in November
11	CNBC	2026-01-30	Trump, two sons, Trump Org sue IRS, Treasury for \$10 billion over tax records leak
12	Reuters	2026-01-31	Nvidia's plan to invest up to \$100 billion in OpenAI has stalled, WSJ reports
13	Reuters	2026-02-02	Trump launches \$12 billion minerals stockpile to boost US manufacturing, counter China
14	CNN	2026-02-04	Bezos-owned storied newspaper Washington Post rolling out mass layoffs
15	ABC News	2026-02-05	US applications for jobless benefits jump to 231,000 last week, the most in 2 months.
16	ABC News	2026-02-06	Pizza Hut closing 250 US stores as parent company considers selling the brand
17	NBC News	2026-02-13	Justice Department's antitrust chief says she's leaving, effective immediately
18	Reuters	2026-02-13	NY Fed report says Americans pay for almost all of Trump's tariffs
19	NBC News	2026-02-20	Supreme Court strikes down most of Trump's tariffs in a major blow to the president
20	NBC News	2026-02-24	FedEx sues Trump administration for tariff refunds after Supreme Court ruling
Society			
1	ABC News	2026-01-06	Epstein files: 2 million records in various stages of review, DOJ says
2	NYT	2026-01-10	Trump Administration Freezes Food Stamps in Minnesota
3	NYT	2026-01-13	Six Prosecutors Quit Over Push to Investigate ICE Shooting Victim's Widow
4	NBC News	2026-01-13	Actor Timothy Busfield turns himself in amid child sex abuse allegations
5	NPR	2026-01-14	Trump administration sends letter wiping out addiction, mental health grants
6	BBC	2026-01-16	Mother of Elon Musk's child sues xAI over Grok deepfakes
7	NBC News	2026-01-16	Immigration officers around Minneapolis are approaching people and demanding proof that they're U.S. citizens
8	NBC News	2026-01-20	Third immigrant detainee at facility in El Paso has died ICE says
9	ABC News	2026-01-22	Uvalde trial: Former school police officer Adrian Gonzales found not guilty on all counts
10	BBC	2026-01-22	ICE detains 5-year-old Liam Ramos during Minnesota operation
11	Reuters	2026-01-28	Meta CEO Zuckerberg blocked curbs on sex-talking chatbots for minors, court filing alleges
12	CNN	2026-01-29	A Texas man detained by ICE was his disabled son's sole caregiver. His son will be laid to rest without him
13	BBC	2026-02-01	Melania director Brett Ratner pictured cuddling woman in Epstein files
14	ABC News	2026-02-04	Attempted Trump assassin sentenced to life in prison
15	NBC News	2026-02-05	9-year-old Illinois boy severely burned after TikTok trend inspired him to microwave NeeDoh toy
16	ABC News	2026-02-06	4 times as many measles cases in a few weeks than US typically averages in a whole year: CDC
17	BBC	2026-02-09	Discord to start requiring face scan or ID to access adult content
18	CNBC	2026-02-10	Moderna says FDA refuses to review its application for experimental flu shot
19	NBC News	2026-02-17	Rev. Jesse Jackson, civil rights icon and two-time presidential candidate, dies at 84
20	NYT	2026-02-20	Labor Secretary's Husband Barred From Department Premises After Reports of Sexual Assaults
Controversy			
1	NPR	2026-01-06	Pentagon will begin review of 'effectiveness' of women in ground combat positions
2	ABC News	2026-01-06	HHS freezes \$10 billion in child care funding for 5 Democratic states, alleging fraudulent programs
3	NYT	2026-01-09	Texas A&M Warns Professor Not to Teach Plato Because of Gender Rules (Gift Article)
4	NYT	2026-01-13	Pentagon used a secret aircraft painted to look like a civilian plane killing 11 people in September
5	Reuters	2026-01-14	Trump says federal payments to sanctuary cities to end from February 1
6	NYT	2026-01-16	Trump Sets Fraudster Free From Prison for a Second Time
7	NBC News	2026-01-22	ICE says its officers can forcibly enter homes during immigration operations without judicial warrants: 2025 memo
8	NYT	2026-01-23	Rejecting Decades of Science, Vaccine Panel Chair Says Polio and Other Shots Should Be Optional
9	Reuters	2026-01-25	Trump moved to cut funding for ICE body cameras, pared back oversight
10	NBC News	2026-01-26	Alex Petti's killing was recorded on body-camera videos, DHS says
11	CNN	2026-01-30	Luigi Mangione will not face death penalty, judge rules
12	CNN	2026-02-01	Leftist and liberal gun groups are seeing a rush of new members
13	Reuters	2026-02-06	Pardoned January 6 rioter pleads guilty to threatening US Democratic leader Jeffries
14	BBC	2026-02-07	Trump removes video with racist clip depicting Obamas as apes
15	ABC News	2026-02-09	'Take the vaccine, please,' a top US health official says in an appeal as measles cases rise
16	BBC	2026-02-10	British Woman shot by dad in Texas after 'arguing about Donald Trump'
17	NBC News	2026-02-10	Trump administration removes Pride flag from Stonewall National Monument, a major LGBTQ landmark
18	NBC News	2026-02-12	EPA reverses longstanding climate change finding, stripping its own ability to regulate emissions
19	ABC News	2026-02-24	Swastika found at Coast Guard training center, investigation underway: USCG
20	CNN	2026-02-25	Anthropic ditches its core safety promise in the middle of an AI red line fight with the Pentagon
Politics			
1	NBC News	2026-01-07	Judge orders Lindsey Halligan to explain why she's still serving as U.S. attorney after previous ruling against her
2	USA Today	2026-01-08	Exclusive: DOGE cuts prompt scramble to feed troops at remote US base
3	NPR	2026-01-10	National Park Service will void passes with stickers over Trump's face
4	CNBC	2026-01-12	Sen. Kelly sues DOD Sec. Hegseth, says he was punished for 'disfavored political speech'
5	NBC News	2026-01-13	Trump administration to end temporary protected status for Somalis
6	Reuters	2026-01-14	US to suspend visa processing for 75 nations, State Department says
7	Reuters	2026-01-15	No. 2 US ICE official Madison Sheahan leaving agency to run for Congress
8	Reuters	2026-01-17	US Justice Department probing Minnesota Governor Walz, other officials, source says
9	Reuters	2026-01-18	Pentagon readies fifteen hundred troops to possibly deploy to MN
10	Reuters	2026-01-21	France asks for a NATO exercise in Greenland, is ready to participate
11	Reuters	2026-01-27	Gregory Bovino removed from his role as US Border Patrol commander at large
12	NBC News	2026-01-27	FBI is investigating Minnesota Signal groups tracking ICE, Patel says
13	NPR	2026-01-28	The Trump administration has secretly rewritten nuclear safety rules
14	BBC	2026-01-31	US government partially shuts down despite funding deal
15	NPR	2026-02-07	State Department will delete X posts from before Trump returned to office
16	ABC News	2026-02-08	Judge orders 3 families returned to US after finding agents used 'lies' to deport them
17	Reuters	2026-02-11	IRS improperly disclosed confidential immigrant tax data to DHS, Washington Post says
18	BBC	2026-02-13	Apple warned by FTC over alleged lack of conservative news
19	Reuters	2026-02-13	Exclusive: US military preparing for potentially weeks-long Iran operations
20	Reuters	2026-02-18	FCC chair denies censoring Colbert interview, confirms investigating ABC's 'The View'

Table 7: All 80 stimulus articles, 20 per topic category, numbered within category and ordered by Reddit post date (column definitions in Appendix E.3).

two measures capture different aspects of diversity: N-matched vocabulary captures the broader type inventory, while MATTR captures within-window repetition. Together they indicate that MACM is not collapsing into stylistic uniformity.

Readability. Flesch–Kincaid grade level shows the largest cross-condition difference: MACM (mean 9.99) reads at roughly 2 grade levels above both Human (7.22) and LGCM (8.06), while LGCM and Human differ by less than half a grade (Cohen’s $d = -0.25$). The proximity of LGCM to Human is intentional and useful: it provides a low-stylistic-shift uncoordinated LLM reference point, while MACM exhibits the more formal register produced by the coordinated multi-agent pipeline.

Stylistic markers. MACM uses notably fewer first-person pronouns per 100 tokens (1.71 vs. 2.35 Human, 2.66 LGCM), consistent with its design of distributing voice across multiple personas rather than concentrating it in a single speaker. All other rate-based markers in Table 8 (question prevalence, exclamation prevalence, all-caps token rate) use the same alphabetic tokenizer ([A-Za-z][A-Za-z’-]*; numbers, URLs, and emoji excluded) so that cells are directly comparable.

Statistical checks. We test whether the three comment pools differ on the two per-comment scalar distributions (words/comment and FK grade) using Welch’s t -test (mean shift), Mann–Whitney U (rank/median), and Cohen’s d (effect size). With N in the tens of thousands, p -values approach zero for any non-zero difference, so the relevant quantity is $|d|$. Selected results:

- Words/comment, Human vs. LGCM: $t = 9.34$, $p < .001$; Mann–Whitney $p = .93$; $d = +0.88$.
- Words/comment, Human vs. MACM: $t = 8.76$, $p < .001$; Mann–Whitney $p = .37$; $d = +0.65$.
- Words/comment, LGCM vs. MACM: $t = -9.46$, $p < .001$; Mann–Whitney $p < .001$; $d = -0.06$.
- FK grade, Human vs. MACM: $t = -24.20$, $p < .001$; $d = -0.70$.
- FK grade, Human vs. LGCM: $t = -7.33$, $p < .001$; $d = -0.25$.

Welch’s t -test on raw words/comment flags Human as significantly longer than either AI condition, but Mann–Whitney U is non-significant for both

Human–LGCM and Human–MACM, and the medians are essentially identical (16 / 15 / 16 words); typical comment length is therefore matched across the three corpora. For FK grade, MACM reads at ≈ 2 grade levels above both Human and LGCM ($d = -0.70$ vs. Human), reflecting the more formal register produced by the coordinated multi-agent pipeline; LGCM is much closer to Human ($d = -0.25$), making it a comparatively low-stylistic-shift uncoordinated LLM reference point.

G Behavioral Replication: Shifts in Participants’ Own Comments

This appendix supplies a supplementary behavioral analysis referenced in §6.

Goal. The body of the paper documents an AI–Human asymmetry on *readers* (RQ1: 32.9% vs. 24.9% change rate), measured through self-reported agreement ratings. To test the same asymmetry without relying on self-report, we ask whether exposure to AI comments also moves what participants *write* themselves.

Sample. $N = 4,996$ within-participant comment pairs. For each article, participants wrote one comment before seeing the comment section and one after, so every pair contrasts the same person’s writing before and after exposure. A 3-LLM ensemble classified each comment as promote, suppress, or neutral, and the change between the two was signed toward the campaign direction.

Result. Participants in the AI conditions moved further toward the campaign direction than those in the Human condition, whose own comments barely shifted ($M = +0.03$): Cohen’s $d = 0.12$ for LGCM (Welch’s t -test, $p = .0004$) and $d = 0.12$ for MACM ($p = .0008$). A mixed model with crossed participant and article random intercepts reproduces the effect ($\beta = +0.067$, $p = .008$ for LGCM; $\beta = +0.075$, $p = .003$ for MACM). The two AI conditions do not differ ($p = .88$).

Caveats. Comment directions are assigned by LLM classifiers rather than human coders, and the same participants supply both this measure and the RQ1 ratings, so the two are consistent with rather than independent of each other.

Ensemble specification. Three LLMs independently classified each comment into {promote, suppress, neutral} using identical prompts

Metric	Human	LGCM	MACM
Comments (N)	1,600	57,600	57,600
Words / comment: mean (sd); median	25.1 (33.5); 16	17.3 (7.1); 15	17.7 (10.1); 16
Sentences / comment (mean)	1.99	1.69	1.72
Vocabulary @ $N=1,600$ (bootstrap mean types)	6,383	4,091	5,144
MATTR (window=100, corpus-level)	0.774	0.783	0.848
Flesch–Kincaid grade: mean; median	7.22; 6.73	8.06; 7.77	9.99; 9.62
Avg. sentence length (words)	12.52	11.30	11.24
First-person pronouns (per 100 tokens)	2.35	2.66	1.71
Question prevalence (% comments with ≥ 1 ?)	21.4%	29.6%	37.9%
Exclamation prevalence (% comments with ≥ 1 !)	6.9%	0.5%	1.9%
All-caps tokens (per 100 tokens)	1.32	0.71	0.87

Table 8: Comment-level corpus statistics for the Human, LGCM, and MACM comment pools (80 articles each). Length numbers are alphabetic-token counts using the regex $[A-Za-z][A-Za-z'-]*$ (numbers, URLs, and emoji excluded). Readability uses Flesch–Kincaid grade level (per-comment, winsorised at $[-5, 30]$, comments with ≥ 5 word tokens). Vocabulary @ $N=1,600$ down-samples each pool to the smallest pool’s N and averages over 50 bootstrap draws to remove the sample-size confound; MATTR uses a 100-token sliding window over the concatenated corpus. First-person pronouns include base forms and apostrophe-preserving contractions (*I’m, we’ve*, etc.); ambiguous collapsed forms (*were, well*) are excluded.

with the article text as context: Claude Sonnet 4.5 (claude-sonnet-4-5), Gemini 3.1 Pro Preview (gemini-3.1-pro-preview), and GPT-5 (gpt-5-2025-08-07). The final label was determined by majority vote ($\geq 2/3$); unanimous disagreements (9.7%) were coded as neutral (Fleiss’ $\kappa = 0.42$).

H Article Tone and Alignment-Stratified Decomposition

H.1 Article tone classification (Sonnet-4 labeller)

Each of the 80 stimulus articles was independently classified by a Claude Sonnet-4 categorical labeler as *positive* / *neutral* / *negative*, using a fixed prompt that asked for the dominant emotional valence of the article body (not the headline alone).

Prevalence. 64 articles (80.0%) were classified *negative*, reflecting the structural negativity bias of *r/news*; 8 *positive* (10.0%) and 8 *neutral* (10.0%). Because of the sparsity of non-negative articles, the primary moderation analysis (§5.5) uses the binary contrast *non-negative* (*pos+neu*, $n = 16$) vs. *negative* ($n = 64$).

Tone $\times$ direction breakdown. Table 9 presents opinion change by the binary tone grouping and campaign direction, pooling across all AI sources and label conditions.

Tone bin	Direction	CR	M
Non-negative	Promote	39.9	+0.49
Non-negative	Suppress	32.7	+0.25
Negative	Promote	31.1	+0.27
Negative	Suppress	33.3	+0.26

Table 9: Article tone $\times$ campaign direction. CR = change rate (%); M = mean opinion change. Non-negative = positive + neutral ($n = 16$ articles).

Non-negative $\times$ Promote is the most effective cell: Welch’s $t = 2.44$, $p = .015$, $d = +0.15$ vs. Negative $\times$ Promote. Suppress direction shows no tone difference ($p = .81$). The Tone $\times$ Direction interaction is marginal ($F = 3.40$, $p = .065$).

H.2 Exploratory pre-exposure moderators

Comment-trust habit. As an exploratory pre-exposure moderator, participants’ baseline trust in online comments ($Q_{\text{Pre},8}$) was the strongest attitudinal predictor of comment evaluation: it correlated with perceived persuasiveness ($r_s = .239$, $p < .0001$) and believability ($r_s = .160$, $p < .0001$). This pattern suggests that habitual trust in comment sections predicts greater receptivity to AI comments rather than resistance to them.

H.3 Alignment-stratified decomposition (24-cell)

This subsection expands the alignment-stratified mechanism map referenced in §5.2, §5.4, and §5.5. All AI observations ($N = 3,330$) were partitioned into 24 cells crossing alignment (SAME, NEUTRAL, OPPOSITE), campaign direction (promote,

suppress), source (LGCM, MACM), and label (No-Label, Label).

Cell definitions.

- **SAME:** baseline Likert is on the same side as the campaign direction (e.g., baseline 5–7 under promote, baseline 1–3 under suppress).
- **NEUTRAL:** baseline = 4.
- **OPPOSITE:** baseline is on the opposite side of the campaign direction.

Headline cells. Successful manipulation is strongest in OPPOSITE cells: under NoLabel, OPPOSITE × Promote produced 53.3% change rate for LGCM ($N = 122$) and 58.2% for MACM ($N = 122$), while OPPOSITE × Suppress produced 45.5% for LGCM ($N = 233$) and 53.4% for MACM ($N = 206$).

- Variance compression by alignment × source.** Under NoLabel, AI-source variance change ($\Delta\text{Var} = (\text{Var}(\text{Treatment}) - \text{Var}(\text{Baseline}))/\text{Var}(\text{Baseline})$; negative = diversity reduction):
- OPPOSITE: LGCM −27.0%, MACM −33.6% (both compress).
 - SAME: LGCM +13.1% (expand), MACM −15.7% (compress).

The LGCM SAME-expansion offsets the LGCM OPPOSITE-compression, leaving aggregate diversity unchanged. MACM compresses on both subgroups, producing the 11.9% aggregate diversity reduction reported in Table 1.

Full 24-cell table. Table 10 reports change rate (%) and mean opinion change for each Alignment × Source × Label × Direction cell.

Align.	Source	Label	Promote		Suppress	
			CR	<i>M</i>	CR	<i>M</i>
SAME	LGCM	NoLabel	26.6	−0.23	14.0	−0.82
	LGCM	Label	15.6	−0.43	11.1	−0.91
	MACM	NoLabel	14.8	−0.65	10.0	−1.27
	MACM	Label	14.9	−0.40	13.6	−0.89
NEUTRAL	LGCM	NoLabel	33.6	+0.37	13.7	−0.34
	LGCM	Label	30.2	+0.28	17.1	−0.12
	MACM	NoLabel	27.5	+0.30	13.8	−0.20
	MACM	Label	38.1	+0.49	18.2	−0.02
OPPOSITE	LGCM	NoLabel	53.3	+1.19	45.5	+0.73
	LGCM	Label	45.7	+0.87	44.9	+0.64
	MACM	NoLabel	58.2	+1.50	53.4	+0.92
	MACM	Label	50.0	+1.15	52.1	+0.97

Table 10: 24-cell decomposition. CR = change rate (%); *M* = mean opinion change.

I Statistical Robustness

This appendix collects the robustness checks referenced throughout §5 and the body of the analysis plan in §4.

I.1 Crossed random-effects LMM specification

Primary specification: $\text{OpinionChange} \sim \text{Source} \times \text{Label} + (1 \mid \text{pid}) + (1 \mid \text{post_id})$, estimated via REML using statsmodels.MixedLM (Seabold and Perktold, 2010). Random intercepts for participant and article account for clustering on both axes.

I.2 Cluster-robust OLS robustness checks

For each focal LMM, an OLS specification $\text{OpinionChange} \sim \text{Source} \times \text{Label}$ with cluster-robust SE was fit twice (clustering by participant and by article). Reported *p*-values bracket the LMM Wald *p* between the two clustering choices. For the aggregate Source × Label interaction (Wald *p* = .066), the bracket is .04–.07.

I.3 Participant-cluster bootstrap

500 participant-resampled iterations of the focal LMM produce a 95% percentile CI of [−0.023, +0.401] for the aggregate Source × Label interaction.

I.4 Permutation test for variance reduction

The variance-reduction test for RQ3 uses 5,000 participant-clustered bootstrap resamples on the $\text{Var}(\text{Treatment})/\text{Var}(\text{Baseline})$ ratio (a ratio < 1

indicates diversity reduction); CIs reported in Table 1 are bias-corrected-and-accelerated (Efron and Tibshirani, 1993). A complementary permutation test that re-shuffles the Source label within participant returns the same significance pattern.

I.5 Full 6-cell breakdown

Table 11 reports the 3×2 Source $\times$ Label cell-level descriptive statistics and pairwise contrasts against the Human+NoLabel baseline.

Source	Label	M [95% CI]	Shift%	d vs. Human
Human	NoLabel	+0.02 [−0.08, +0.12]	26.0%	N/A
Human	Label	−0.03 [−0.13, +0.07]	23.7%	N/A
LGCM	NoLabel	+0.29 [+0.19, +0.39]	34.0%	0.19 ($p < .001$)
LGCM	Label	+0.17 [+0.08, +0.26]	29.7%	0.14 ($p = .004$)
MACM	NoLabel	+0.27 [+0.16, +0.37]	32.5%	0.16 ($p < .001$)
MACM	Label	+0.42 [+0.32, +0.52]	35.6%	0.31 ($p < .001$)

Table 11: 3×2 Source $\times$ Label cell descriptives ($N \approx 830$ –840/cell). M = mean opinion change; 95% CI from 5,000 participant-clustered bootstrap; Shift% = share with opinion change > 0 ; d = Cohen’s d vs. Human+NoLabel. SD ranges from 1.37 to 1.57 across cells. Cell-wise medians are uniformly zero (Appendix I.6), so mean and bootstrap CI rather than median [IQR] are used for between-cell comparison.

The 3×2 ANOVA confirms a significant Source main effect ($F(2, 4994) = 25.20, p < .0001, \eta_p^2 = .010$), no Label main effect ($F(1, 4994) = 0.02, p = .882$), and a significant Source $\times$ Label interaction ($F(2, 4994) = 4.07, p = .017$): labeling reduces LGCM persuasion ($+0.29 \rightarrow +0.17$) but increases MACM persuasion ($+0.27 \rightarrow +0.42$). A nonparametric Kruskal–Wallis test across the six cells reproduces the source main effect ($H = 57.66, df = 5, p < .001$); per-cell nonparametric robustness checks are reported in Appendix I.7.

I.6 Distribution of opinion change

Opinion change is the difference of two 7-point Likert items and takes 13 integer values (−6 to +6). Table 12 reports the per-cell frequency of each value. Across the six cells, 44% to 51% of observations are zero. Consequently, cell-wise medians are uniformly zero, with first and third quartiles of 0 and +1, respectively. This motivates our choice of mean with participant-clustered bootstrap 95%

CI (Table 11) over median [IQR] for between-cell comparison; the cell-wise medians are completely uninformative for this dataset.

Opinion change	Human		LGCM		MACM	
	NoLabel	Label	NoLabel	Label	NoLabel	Label
−6	0.4%	0.5%	0.0%	0.0%	0.0%	0.0%
−5	0.5%	0.6%	0.1%	0.0%	0.0%	0.1%
−4	0.7%	0.5%	0.8%	1.0%	1.2%	0.5%
−3	2.6%	2.3%	1.7%	2.0%	1.9%	1.4%
−2	7.3%	6.9%	5.9%	6.1%	6.7%	4.3%
−1	13.1%	14.7%	10.4%	12.3%	13.6%	10.4%
0	49.5%	50.8%	47.1%	48.9%	44.1%	47.7%
+1	15.9%	14.7%	18.3%	16.2%	15.6%	17.6%
+2	5.3%	5.4%	9.0%	8.7%	7.8%	9.4%
+3	2.4%	1.8%	4.1%	2.8%	5.9%	5.2%
+4	1.2%	0.7%	1.6%	1.2%	1.5%	1.8%
+5	0.7%	0.2%	0.6%	0.5%	0.8%	0.8%
+6	0.5%	0.8%	0.4%	0.4%	0.8%	0.7%
Median	0	0	0	0	0	0
IQR	1	1	1	1	1	1

Table 12: Per-cell frequency distribution of opinion change. Zero accounts for 44% to 51% of observations across cells; cell-wise medians and IQRs are uniformly 0 and 1, respectively, and do not distinguish between conditions.

I.7 Normality and nonparametric robustness

Because opinion change is ordinal with a large mass at zero (Appendix I.6), we verify every parametric inference in §5 with its nonparametric counterpart. The cell-level normality assessment (Table 13), pairwise nonparametric comparisons (Table 14), and within-cell one-sample tests against zero (Table 15) jointly confirm that all conclusions reported in §5 hold under nonparametric inference.

Cell	N	W	p	Normal?
Human-NoLabel	835	0.875	$< .001$	No
Human-Label	835	0.854	$< .001$	No
LGCM-NoLabel	830	0.902	$< .001$	No
LGCM-Label	835	0.894	$< .001$	No
MACM-NoLabel	825	0.909	$< .001$	No
MACM-Label	840	0.888	$< .001$	No

Table 13: Shapiro–Wilk normality test on opinion change by cell. All six cells reject the normality null, motivating the nonparametric robustness checks below.

Pair	M_1	M_2	Welch t	p_t	Mann–Whitney U	p_U	d	r
LGCM-NoLab vs. MACM-NoLab	+0.29	+0.27	+0.32	.747	352,968	.250	−0.02	−0.03
LGCM-Label vs. MACM-Label	+0.17	+0.42	−3.66	< .001	319,263	< .001	+0.18	+0.09
LGCM-NoLab vs. LGCM-Label	+0.29	+0.17	+1.78	.076	364,594	.050	−0.09	−0.05
MACM-NoLab vs. MACM-Label	+0.27	+0.42	−2.09	.037	323,056	.012	+0.10	+0.07
Human-NoLab vs. LGCM-NoLab	+0.02	+0.29	−3.81	< .001	308,902	< .001	+0.19	+0.11
Human-NoLab vs. MACM-NoLab	+0.02	+0.27	−3.30	< .001	319,462	.007	+0.16	+0.07
Human-NoLab vs. Human-Label	+0.02	−0.03	+0.73	.467	356,451	.393	−0.04	−0.02

Table 14: Pairwise focal comparisons on opinion change: Welch’s t (parametric, Cohen’s d) and Mann–Whitney U (nonparametric, rank-biserial r). Bold rows indicate $p < .05$ under *both* tests. The LGCM-NoLab vs. LGCM-Label contrast is marginal in both directions ($p_t = .076$, $p_U = .050$). For all bold (significant) rows, parametric and nonparametric inferences agree.

Cell	N	M	Wilcoxon W	p (two-sided)
Human-NoLabel	835	+0.02	44,624	.999
Human-Label	835	−0.03	40,065	.332
LGCM-NoLabel	830	+0.29	33,790	< .001
LGCM-Label	835	+0.17	37,428	< .001
MACM-NoLabel	825	+0.27	40,760	< .001
MACM-Label	840	+0.42	27,450	< .001

Table 15: One-sample Wilcoxon signed-rank test of opinion change versus zero, per cell. Each article-level opinion change value is itself a within-subject paired difference (treatment minus baseline), so the one-sample test against zero is the nonparametric analog of the paired t -test for “campaign-direction shift $\neq 0$ ”. All four AI cells reject the zero-shift null at $p < .001$; both Human cells do not reject ($p > .30$), confirming Human comments as a natural baseline.

Summary. The nonparametric tests reproduce every parametric conclusion reported in §5: the Source main effect (Kruskal–Wallis $H = 57.66$, $p < .001$), the LGCM/MACM equivalence on the whole sample (Mann–Whitney $p = .250$ in No-Label), the within-cell shift for all four AI cells (Wilcoxon $p < .001$) and the null shift for both Human cells (Wilcoxon $p > .30$). The mean with participant-clustered bootstrap 95% CI is used as the primary location measure throughout the paper both because cell-wise medians are uniformly zero (Appendix I.6) and because the parametric and nonparametric inferences agree on every focal comparison.

J Survey Instruments

This appendix documents the survey instruments used in the four study phases (§4). A total of 102 items were administered in the Label condition (95 in the NoLabel condition). All Likert scales used 7 points unless otherwise noted. Questions are indexed as $Q_{\text{Phase}.n}$: **Pre** = Phase 1 pre-survey, **Base** = Phase 2 baseline, **Treat** = Phase 3 treatment,

Post = Phase 4 post-survey.

Provenance of items. Our instrument adapts the measurement approach of prior AI-persuasion studies (Salvi et al., 2025; Bai et al., 2025; Potter et al., 2024) to a news-comment setting. The pre-post agreement items behind our primary dependent variable ($Q_{\text{Base}.2}$, $Q_{\text{Treat}.1}$) reuse their single-proposition Likert design; the comment-evaluation items ($Q_{\text{Treat}.4}$ – $Q_{\text{Treat}.6}$) adapt it so that participants rate a 20-comment thread once per article rather than a single message. Items with no prior counterpart were written for this study: manipulation and coordination awareness ($Q_{\text{Post}.15}$ – $Q_{\text{Post}.20}$), AI-comment detection ($Q_{\text{Post}.26}$ – $Q_{\text{Post}.27}$), and the Label-only items ($Q_{\text{Post}.28}$ – $Q_{\text{Post}.30e}$). Remaining items follow standard survey practice.

J.1 Phase 1: Pre-survey (16 items)

The pre-survey was administered before any article or comment exposure. It measured participants’ baseline news consumption habits ($Q_{\text{Pre}.1}$ – $Q_{\text{Pre}.7}$), comment and article trust ($Q_{\text{Pre}.8}$, $Q_{\text{Pre}.12}$), platform usage ($Q_{\text{Pre}.9}$ – $Q_{\text{Pre}.11}$), political awareness ($Q_{\text{Pre}.13}$ – $Q_{\text{Pre}.14}$), political orientation ($Q_{\text{Pre}.15}$), and included two quality checks: an emotional trap (EM-1, $Q_{\text{Pre}.5}$) and an AI self-report check (AI-1, $Q_{\text{Pre}.16}$). Table 16 lists all 16 items.

J.2 Phase 2: Article reading, Baseline (3 items $\times$ 5 articles)

For each of the 5 articles, participants answered the following **before seeing any comments** (Table 17). $Q_{\text{Base}.2}$ provides the baseline agreement rating ($\text{Agr}_{\text{baseline}}$) used to compute the primary DV.

News & Comment Habits ($Q_{Pre.1}$ – $Q_{Pre.12}$)	
$Q_{Pre.1}$	[News frequency] How often do you read or watch the news? <i>Radio (5): Rarely / A few times a month / A few times a week / Almost every day / Multiple times a day</i>
$Q_{Pre.2}$	[Reading depth] When you read news, how much do you typically read? <i>Radio (4): Headlines only / Headlines + first few paragraphs / Most of the article / Full article</i>
$Q_{Pre.3}$	[Comment habits] When you read news, how many comments do you typically read? <i>Radio (6): None / 1–2 / 3–5 / 6–10 / More than 10 / All or most</i>
$Q_{Pre.4}$	[Comment engagement] How often do you write comments on news articles? <i>Radio (5): Never / Rarely / Sometimes / Often / Very often</i>
$Q_{Pre.5}$	[EM-1 attention check] How are you feeling right now? <i>Radio (5): Tired / Bored / Fine / Stressed / “I do not experience emotions” (trap)</i>
$Q_{Pre.6}$	[Vote behavior] How often do you upvote or downvote comments? <i>Radio (5): Never / Rarely / Sometimes / Often / Very often</i>
$Q_{Pre.7}$	[Reply behavior] How often do you reply to other people’s comments? <i>Radio (5): Never / Rarely / Sometimes / Often / Very often</i>
$Q_{Pre.8}$	[Comment trust, pre] How much do you trust news comments? <i>Trust Likert (7): Strongly distrust / Distrust / Slightly distrust / Neutral / Slightly trust / Trust / Strongly trust</i>
$Q_{Pre.9}$	[Platform use] Where do you primarily read news and comments? (Select all) <i>Checkbox (8): Reddit / Facebook / X / YouTube / TikTok / Instagram / News websites / Other</i>
$Q_{Pre.10}$	[SNS usage] How much time do you spend on social media per day? <i>Radio (5): Almost none / <30 min / 30 min–1 hr / 1–2 hrs / >2 hrs</i>
$Q_{Pre.11}$	[Sort preference] Which comment sorting order do you usually prefer? <i>Radio (5): Default / Top/Best / Newest / Oldest / Not sure</i>
$Q_{Pre.12}$	[Article trust, pre] How much do you trust news article content? <i>Trust Likert (7): Strongly distrust / Distrust / Slightly distrust / Neutral / Slightly trust / Trust / Strongly trust</i>
Political Awareness ($Q_{Pre.13}$ – $Q_{Pre.16}$)	
$Q_{Pre.13}$	[Political interest] How often do you follow political or election news? <i>Radio (5): Rarely / A few times a month / A few times a week / Almost every day / Multiple times a day</i>
$Q_{Pre.14}$	[Political discussion] How often do you discuss politics with others? <i>Radio (5): Rarely / A few times a month / A few times a week / Almost every day / Multiple times a day</i>
$Q_{Pre.15}$	[Political orientation] Where would you place yourself on the political spectrum? <i>Political Likert (7): Very conservative / Conservative / Slightly conservative / Moderate / Slightly liberal / Liberal / Very liberal</i>
$Q_{Pre.16}$	[AI-1 attention check] Did you use any AI tools (ChatGPT, Claude, etc.) to assist with your responses? <i>Radio (2): Yes / No</i>

Table 16: Phase 1: Pre-survey instrument (16 items).

$Q_{Base.1}$	[Reading depth] How much of this article did you read? <i>Radio (4): Headlines only / Headlines + first few paragraphs / Most of it / Full article</i>
$Q_{Base.2}$	[Baseline Agr] How much do you agree with this article’s content? <i>Agreement Likert (7): Strongly disagree / Disagree / Slightly disagree / Neutral / Slightly agree / Agree / Strongly agree</i>
$Q_{Base.3}$	[Opinion strength] How strong is your opinion on this topic? <i>Strength Likert (7): Very weak / Weak / Slightly weak / Moderate / Slightly strong / Strong / Very strong</i>

Table 17: Phase 2: Article reading items (Baseline, 3 items $\times$ 5 articles).

J.3 Phase 3: Comment exposure, Treatment (6 items $\times$ 5 articles)

For each of the 5 articles, participants answered the following **after reading 20 condition-assigned comments** (Table 18). $Q_{Treat.1}$ provides the treatment agreement rating ($Agr_{treatment}$); $Q_{Treat.4}$ – $Q_{Treat.6}$ measure perceived persuasiveness, believability, and quality of the comments.

J.4 Phase 4: Post-survey (34 common + 7 Label-only items)

The post-survey was administered after all 5 articles were completed. It covered study experience

($Q_{Post.1}$ – $Q_{Post.3}$), comment perception and influence ($Q_{Post.4}$ – $Q_{Post.13}$), an attention check (EM-2, $Q_{Post.14}$), fake news and manipulation awareness ($Q_{Post.15}$ – $Q_{Post.20}$), AI attitudes and usage ($Q_{Post.21}$ – $Q_{Post.25}$, including attention check AC-3 at $Q_{Post.25}$), AI comment detection ($Q_{Post.26}$ – $Q_{Post.27}$), and demographics ($Q_{Post.31}$ – $Q_{Post.37}$, including AI self-report check AI-2 at $Q_{Post.36}$). Items $Q_{Post.28}$ – $Q_{Post.30e}$ were administered only in the Label condition. Tables 19 and 20 list all items.

$Q_{\text{Treat.1}}$	[Treatment Agr] How much do you agree with this article's content? <i>Agreement Likert (7): Strongly disagree / Disagree / Slightly disagree / Neutral / Slightly agree / Agree / Strongly agree</i>
$Q_{\text{Treat.2}}$	[Opinion strength] How strong is your opinion on this topic? <i>Strength Likert (7): Very weak / Weak / Slightly weak / Moderate / Slightly strong / Strong / Very strong</i>
$Q_{\text{Treat.3}}$	[Engagement] How many comments did you read for this article? <i>Radio (6): None / 1–2 / 3–5 / 6–10 / More than 10 / All or most</i>
$Q_{\text{Treat.4}}$	[Persuasiveness] How persuasive did you find the comments? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Treat.5}}$	[Believability] How believable did you find the comments? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Treat.6}}$	[Quality] How would you rate the overall quality of the comments? <i>Quality Likert (7): Very low / Low / Slightly low / Average / Slightly high / High / Very high</i>

Table 18: Phase 3: Comment exposure items (Treatment, 6 items $\times$ 5 articles).

Study Experience ($Q_{\text{Post.1}}-Q_{\text{Post.3}}$)	
$Q_{\text{Post.1}}$	[Reading check] How much of the news articles did you typically read during this study? <i>Radio (4): Headlines only / Headlines + first few paragraphs / Most of the article / Full article</i>
$Q_{\text{Post.2}}$	[Comment check] On average, how many comments did you read per article? <i>Radio (6): None / 1–2 / 3–5 / 6–10 / More than 10 / All or most</i>
$Q_{\text{Post.3}}$	[Change pathway] If your opinion changed, what influenced the change? (Select all) <i>Checkbox (6): Did not change / Convincing arguments / Same opinion / Upvotes / Re-reading article / Other</i>
Comment Perception ($Q_{\text{Post.4}}-Q_{\text{Post.13}}$)	
$Q_{\text{Post.4}}$	[Influence belief, pre] Before this study, to what extent did you believe comments could influence your opinion? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Post.5}}$	[Influence belief, post] Now, to what extent do you believe comments can influence your opinion? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Post.6}}$	[Conformity, pre] Before this study, to what extent did you align with majority opinion? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Post.7}}$	[Conformity, post] Now, to what extent do you align with majority opinion? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Post.8}}$	[Comment trust, post] How much do you trust news comments? <i>Trust Likert (7): Strongly distrust / Distrust / Slightly distrust / Neutral / Slightly trust / Trust / Strongly trust</i>
$Q_{\text{Post.9}}$	[Article trust, post] How much do you trust news article content? <i>Trust Likert (7): Strongly distrust / Distrust / Slightly distrust / Neutral / Slightly trust / Trust / Strongly trust</i>
$Q_{\text{Post.10}}$	[Persuasiveness] How persuasive were the comments overall? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Post.11}}$	[Manipulation detect] To what extent did comments try to change your opinion? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Post.12}}$	[Direction detect] To what extent did comments push you toward a particular opinion? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Post.13}}$	[Naturalness] How natural did the comment discussions feel? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Post.13b}}$	[AI detection, qual] Did you notice anything unusual about the comments? <i>Open-ended response (optional free text; no fixed options)</i>
Attention Check ($Q_{\text{Post.14}}$)	
$Q_{\text{Post.14}}$	[EM-2 attention check] Did anything annoying happen to you today? <i>Radio (4): Yes minor / Yes major / No / “I do not experience annoyance” (trap)</i>
Fake News & Manipulation ($Q_{\text{Post.15}}-Q_{\text{Post.20}}$)	
$Q_{\text{Post.15}}$	[Fake news perception] How many articles felt like fake news? <i>Radio (5): None / A few / Some / Many / Almost all</i>
$Q_{\text{Post.16}}$	[Fake news confidence] How confident in telling fake from real news? <i>Confidence Likert (7): Not at all / Not very / A little / Neutral / Somewhat / Fairly / Very</i>
$Q_{\text{Post.17}}$	[AI detection confidence] How confident in identifying AI-generated fake news? <i>Confidence Likert (7): Not at all / Not very / A little / Neutral / Somewhat / Fairly / Very</i>
$Q_{\text{Post.18}}$	[Manipulation awareness] To what extent do you believe online opinion manipulation exists? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Post.19}}$	[Manipulation detection] How many comments seemed part of opinion manipulation? <i>Radio (5): None / A few / Some / Many / Almost all</i>
$Q_{\text{Post.20}}$	[Threat model] Who is behind online manipulation? (Select all) <i>Checkbox (10): Political parties / Corporations / PR firms / Foreign govts / Communities / Influencers / Simple bots / Advanced AI systems / Not sure / Other</i>

Table 19: Phase 4: Post-survey instrument, Part A ($Q_{\text{Post.1}}-Q_{\text{Post.20}}$).

AI Attitudes ($Q_{\text{Post.21}}-Q_{\text{Post.25}}$)	
$Q_{\text{Post.21}}$	[AI affect] How do you feel about AI? <i>Sentiment Likert (7): Very negative / Negative / Slightly negative / Neutral / Slightly positive / Positive / Very positive</i>
$Q_{\text{Post.22}}$	[AI frequency] How often do you use AI tools? <i>Radio (5): Never / Rarely / Sometimes / Often / Very often</i>
$Q_{\text{Post.23}}$	[AI tool usage] Which AI tools do you use? (Select all) <i>Checkbox (6): None / ChatGPT / Claude / Copilot / Gemini / Other</i>
$Q_{\text{Post.24}}$	[AI purpose] What do you mainly use AI for? (Select all) <i>Checkbox (8): Don't use / Info search / Writing / Coding / Translation / Automation / Learning / Other</i>
$Q_{\text{Post.25}}$	[AC-3 attention check] Attention check: select "Neutral." <i>Degree Likert (7): correct answer = 4 (Neutral)</i>
AI Comment Detection ($Q_{\text{Post.26}}-Q_{\text{Post.27}}$)	
$Q_{\text{Post.26}}$	[AI comment confidence] How confident in telling AI comments from human comments? <i>Confidence Likert (7): Not at all / Not very / A little / Neutral / Somewhat / Fairly / Very</i>
$Q_{\text{Post.27}}$	[AI detection, self] Did any comments seem AI-written? <i>Radio (4): None / A few / Quite a few / Most</i>
AI Label Section ($Q_{\text{Post.28}}-Q_{\text{Post.30e}}$; Label condition only)	
$Q_{\text{Post.28}}$	[Label awareness] Did you notice the "Likely AI-generated" labels? <i>Radio (3): Yes / No / Not sure</i>
$Q_{\text{Post.29}}$	[Label → trust] To what extent did the label reduce your trust in those comments? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Post.29b}}$	[Label → trust matrix] After seeing the labels, how did your trust change for each of the following? <i>Likert matrix (5): Strongly decreased / Decreased / No change / Increased / Strongly increased</i> <i>Rows: (a) Comments with the AI label, (b) Comments without the AI label, (c) The news articles themselves, (d) Online comments in general</i>
$Q_{\text{Post.30}}$	[Label → opinion] To what extent did the label influence your opinion formation? <i>Degree Likert (7): Not at all / Very little / A little / Neutral / Somewhat / Quite a bit / Very much</i>
$Q_{\text{Post.30b}}$	[Label mechanism, qual] How did the label affect how you read comments? <i>Open-ended response (optional free text; no fixed options)</i>
$Q_{\text{Post.30c}}$	[Policy preference] Should platforms indicate AI-generated comments? <i>Radio (3): Yes / No / Not sure</i>
$Q_{\text{Post.30d}}$	[Display preference] How should it be displayed? (if $Q_{\text{Post.30c}}$ =Yes) <i>Radio (5): Label each / Label + probability / Summary only / Both / Other</i>
$Q_{\text{Post.30e}}$	[Policy rationale] Why / why not? (based on $Q_{\text{Post.30c}}$) <i>Checkbox: Yes (6 reasons) / No (7 reasons)</i>
Demographics & Final Checks ($Q_{\text{Post.31}}-Q_{\text{Post.37}}$)	
$Q_{\text{Post.31}}$	[Demographics] Age <i>Select: 18–99+</i>
$Q_{\text{Post.32}}$	[Demographics] Gender <i>Radio (4): Male / Female / Non-binary / Prefer not to say</i>
$Q_{\text{Post.33}}$	[Demographics] Highest education completed <i>Radio (8): < High school / High school or GED / Some college / Associate's / Bachelor's / Master's / Doctoral / Professional degree</i>
$Q_{\text{Post.34}}$	[Demographics] Field of study or interest (Select all) <i>Checkbox (8): Humanities / Social Sci / Natural Sci / Engineering / Business / Health / Education / Other</i>
$Q_{\text{Post.35}}$	[Political shift] Where would you place yourself on the political spectrum? <i>Political Likert (7): Very conservative / Conservative / Slightly conservative / Moderate / Slightly liberal / Liberal / Very liberal</i>
$Q_{\text{Post.36}}$	[AI-2 attention check] Did you use any AI tools during this survey? <i>Radio (2): Yes / No</i>
$Q_{\text{Post.37}}$	[Free feedback] Any thoughts or feedback about this study? <i>Open-ended response (optional free text; no fixed options)</i>

Table 20: Phase 4: Post-survey instrument, Part B ($Q_{\text{Post.21}}-Q_{\text{Post.37}}$). Items $Q_{\text{Post.28}}-Q_{\text{Post.30e}}$ were administered only in the Label condition.

K Study Environment Screenshots

This appendix shows the custom web application used in the experiment, which replicates Reddit's r/news interface.

L Participant Demographics

Table 21 summarizes the demographic composition of the final analytic sample ($N = 1,000$). Table 22 confirms that randomization produced balanced cells across all six conditions. Participants were compensated \$6 for the 30-minute session

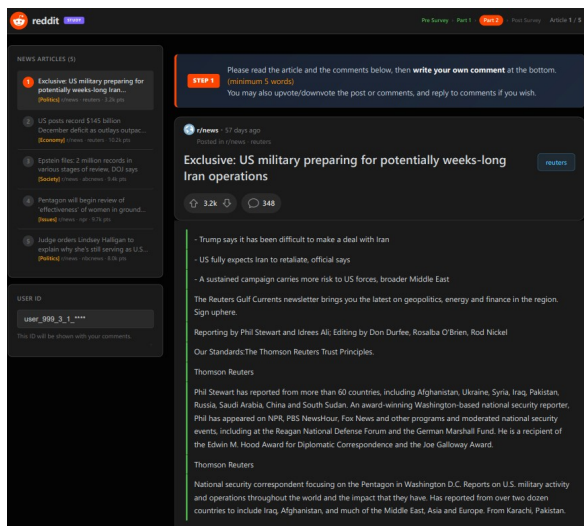


Figure 3: Overview of the study environment. Participants read a news article on a Reddit-like platform; comments appeared below the article body. Left sidebar shows the 5 assigned articles.

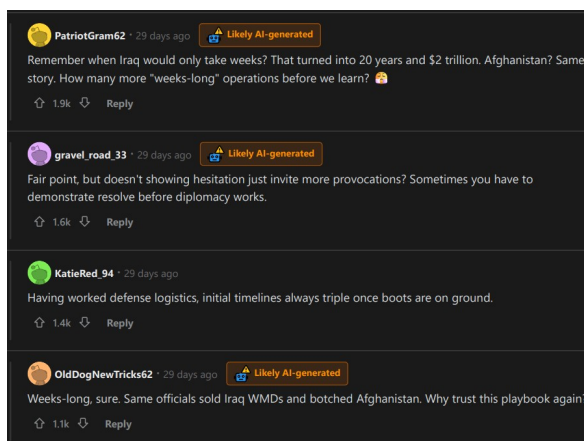


Figure 4: The Label condition with MACM-generated comments. Comments flagged by XDAC display a “Likely AI-generated” badge (orange). Unflagged comments (e.g., KatieRed_94) evaded detection and appear without the badge.

(\$12/hr), following CloudResearch’s recommended rate guidelines.

Variable	Category	%
Gender	Female	49.5
	Male	48.4
Age	18–29	16.7
	30–39	30.5
	40–49	25.7
	50–59	14.6
	60+	10.4
Political party	Republican	29.9
	Democrat	29.5
	Independent	38.5
Race/Ethnicity	White	74.7
	Black / African American	12.3
	Hispanic / Latino	9.5
	Asian / Pacific Islander	6.7
	Other	4.2
Education	High school or less	13.5
	Some college / Associate	33.8
	Bachelor’s	35.4
	Graduate+	15.2
Employment	Full-time	46.4
	Part-time	13.5
	Not in paid work	38.0

Table 21: Demographic composition of the final sample ($N = 1,000$). Political party and gender reflect quota targets; remaining variables were drawn via Census Match. Age: $M = 41.8$, $SD = 12.9$, range 18–82. Household income: $Mdn = \$65K$, $M = \$79K$ ($SD = \$57K$). Race/Ethnicity categories are not mutually exclusive (Hispanic/Latino is an ethnicity overlay); “Other” includes American Indian/Alaska Native ($n = 14$) and unlisted ($n = 28$). 21 participants lacked CloudResearch profile linkage; percentages are computed over respondents.

Type	Variable	Statistic	<i>p</i>
Demo.	Age	$F(5, 973) = 1.84$	.102
	Gender	$\chi^2(10) = 12.11$	.278
	Political party	$\chi^2(15) = 15.64$	.407
	Race	$\chi^2(25) = 23.83$	.529
	Hispanic/Latino	$\chi^2(10) = 12.93$	.228
	Education	$\chi^2(20) = 31.65$	.047
	Employment	$\chi^2(10) = 9.52$	.484
	Income	$F(5, 882) = 1.79$	.111
Pre-s.	Comment trust	$F(5, 989) = 0.47$	.801
	Article trust	$F(5, 989) = 0.53$	.757
	AI detection conf.	$F(5, 989) = 0.53$	.757
	Prior AI tool use	$\chi^2(5) = 5.82$	.324
	Political leaning (Q15)	$F(5, 989) = 0.53$	.756
	AI sentiment (Q21)	$F(5, 994) = 0.61$	.691
	AI use freq. (Q22)	$F(5, 994) = 0.77$	.574

Table 22: Randomization balance across six conditions (3 Source $\times$ 2 Label). ANOVA (F) for continuous variables; χ^2 for categorical variables. All variables show comparable distributions across conditions; Education is marginal ($p = .047$) but substantively negligible. All pre-survey measures, including political leaning and two AI familiarity items, show balanced baseline attitudes (all $p > .30$).

M User Attitudes Toward AI Labeling

In the post-survey (Phase 4; Appendix J), Label-condition participants ($n = 502$) answered three additional questions about whether platforms should indicate AI-generated comments ($Q_{\text{Post.30c}}$), how such indicators should be displayed ($Q_{\text{Post.30d}}$), and why they support or oppose labeling ($Q_{\text{Post.30e}}$). Table 23 summarizes the first two.

Response	<i>n</i>	%
$Q_{\text{Post.30c}}$: Should platforms indicate AI-generated comments?		
Yes	424	84.5%
No	20	4.0%
Not sure	58	11.6%
$Q_{\text{Post.30d}}$: How should it be displayed?		
Label on each suspected comment	179	42.4%
Label + AI probability on each	165	39.1%
Both summary & individual labels	62	14.7%
Summary at top only	14	3.3%
Other	2	0.5%

Table 23: Label-condition participants’ policy preferences regarding AI-comment disclosure. $Q_{\text{Post.30c}}$ was administered to all 502 Label-condition participants; $Q_{\text{Post.30d}}$ summarizes display preferences among those who answered *Yes*.

Among the 424 respondents who favored labeling ($Q_{\text{Post.30e}}$, multiple choice), the most cited reasons were: *readers deserve to know the source*

(80.1%), *helps evaluate critically* (71.6%), *reduce misinformation* (70.9%), *protect public discourse* (55.6%), and *minimal awareness for self-protection* (55.1%). The 20 respondents who opposed labeling ($Q_{\text{Post.30e}}$) most frequently cited: *detection not accurate enough* (70.0%), *could be used to censor viewpoints* (60.0%), and *content should be judged by merit, not source* (30.0%).

These preferences were consistent across comment sources (Human 86.2%, LGCM 80.8%, MACM 86.3% in favor) and across label-awareness levels ($Q_{\text{Post.28}}$: noticed 84.8%, did not notice 83.1% in favor), suggesting that the demand for labeling reflects a general norm rather than a reaction to observing specific labels.

N Label Comprehension and Behavioral Response

A natural question is whether the label’s failure to reduce persuasion reflects a failure of *comprehension*, meaning that participants may not have understood what the label meant. To test this, we examined two post-survey trust measures (Phase 4; see Table 20 in Appendix J). $Q_{\text{Post.29b}}$ asked Label-condition participants how the label changed their trust in labeled comments (5-point scale: 1=strongly decreased to 5=strongly increased). Among all 502 participants, 69% reported decreased trust (ratings 1–2; $M = 1.99$, $SD = 0.95$), 26% reported no change, and only 6% reported increased trust. A separate item, $Q_{\text{Post.29}}$, measured the overall extent of trust reduction on a 7-point scale (1=not at all to 7=very much). Label comprehension was thus high: most participants understood the label as a negative credibility signal.

We then tested whether this comprehension translated into behavioral resistance. Among Label-condition AI participants (LGCM + MACM; $n = 335$), we split by $Q_{\text{Post.29}}$ response. Participants reporting high trust reduction (≥ 5 ; $n = 233$; mean opinion change = +0.099) were not significantly less persuaded than those reporting minimal reduction (≤ 3 ; $n = 61$; mean opinion change = +0.030; Welch’s $t = 0.57$, $p = .57$, $d = 0.08$; Mann–Whitney $p = .93$). The Spearman correlation between $Q_{\text{Post.29}}$ and opinion change was negligible ($r_s = -0.03$, $p = .62$). Comprehension alone did not ensure resistance.

Open-ended responses to $Q_{\text{Post.30b}}$ (“How did the label affect how you read the comments?”; $n = 335$ responses among AI-Label participants) illustrate

the dissociation. Below, Q_{29} denotes self-reported trust reduction (7-point; 1=not at all, 7=very much), and opinion change is the per-participant mean opinion change across five articles.

Persuaded despite comprehension. Among participants who shifted most toward the campaign direction despite reporting reduced trust:

- “*It made me ‘down play’ them in my mind.*” (MACM participant, $Q_{29} = 5$ [somewhat], opinion change = +3.0)
- “*That it was simply a generated opinion with no one behind it, it was empty. Just used to let people feel validated.*” (LGCM participant, $Q_{29} = 6$ [quite a bit], opinion change = +2.4)

Both participants recognized the AI source and reported reduced trust, yet their opinions shifted substantially toward the campaign direction.

Resisted via disengagement. By contrast, among participants who shifted most against the campaign direction and actively disengaged:

- “*I passed over most of them because I didn’t think they were authentic.*” (MACM participant, $Q_{29} = 7$ [very much], opinion change = −2.6)
- “*I didn’t believe them. I took them with a grain of salt.*” (MACM participant, $Q_{29} = 6$ [quite a bit], opinion change = −2.2)

The minority who resisted did so by actively disengaging from labeled content, a behavioral response that most comprehending participants did not adopt.

O Mitigation Framework

Our RQ4 results demonstrate that per-comment AI labels, the sole intervention currently mandated or proposed by major transparency regulations (EU AI Act Article 50, U.S. transparency bills), mitigate persuasion in only 1 of 4 Source × Direction cells, while amplifying it in another (§5.4). This asymmetry motivates a layered defense: no single mechanism addresses the full threat surface revealed by RQ1–RQ4. Table 24 outlines four complementary layers, ordered by scope.

L1: Per-comment AI label (tested). This is the only layer directly evaluated in the present RCT. Labels raised source recognition nearly threefold (d' : 0.36 → 1.07) and 69% of participants reported decreased trust in labeled comments (Appendix N), yet behavioral persuasion was unchanged in aggregate (33.2% → 32.7%; participant-clustered GEE, OR = 0.97, 95% CI [0.83, 1.15], $p = .76$). The

label is necessary but not sufficient: it produces no false-positive harm and satisfies users’ strong preference for disclosure (84.5% support; Appendix M), yet it leaves behavioral persuasion unchanged.

L2: Exposure restriction and social-proof stripping (proposed). RQ1 shows that even a single exposure session converts 32.9% of AI-exposed readers. L2 targets the *quantity* of exposure: flagged comments are collapsed, down-ranked, stripped of upvote/like counts, or segregated into a separate view. These mechanisms mirror infrastructure platforms already deploy for spam and abuse filtering, requiring no new detection capabilities beyond those powering L1.

L3: Campaign-level warning banner (proposed). RQ3 reveals that coordinated AI compresses opinion diversity at the *section* level, a phenomenon invisible to per-comment cues. L3 addresses this gap by displaying a banner when a comment section shows anomalously uniform directionality (e.g., “This comment section shows unusually uniform directionality and may include AI-coordinated activity”). Unlike L1, which flags individual items, L3 communicates the *structural* anomaly that coordination produces.

L4: Diversity injection (proposed). RQ3’s variance decomposition shows that MACM suppresses visible dissent, manufacturing apparent consensus. L4 counters this by *reordering* existing dissenting comments to the top of the thread, surfacing minority voices rather than generating new content. This is grounded in Asch (1956)’s finding that conformity collapses when even one visible dissenter is present, and avoids the editorial-neutrality concerns that content generation would raise.

Complementarity and limitations. The four layers are designed to be complementary: L1 operates at the comment level (credibility), L2 at the exposure level (visibility), L3 at the section level (structural awareness), and L4 at the discourse level (diversity restoration). Only L1 has been empirically tested in this study; L2–L4 are motivated by specific findings but require independent validation. Furthermore, all four layers act on the *reader-facing surface* of a coordinated campaign. Complementary operator-side mechanisms, such as account-level coordination analysis and persistent attribution that raise the cost of large-scale operations, fall outside the scope of this user study

Layer	Mitigation mechanism	Description	Motivated by	Empirical status
L1	Per-comment AI label	Display “Likely AI-generated” badge on individual comments flagged by a detector.	RQ4 (§5.4): labels reduce believability but mitigate persuasion in only 1 of 4 cells	Tested in this study ; necessary but insufficient
L2	Exposure restriction & social-proof stripping	Comments flagged by an AI-detector are <i>collapsed, down-ranked, stripped of social-proof cues</i> (e.g., hidden up-vote/like counts), or <i>segregated into a separate AI-comment view</i> , using the same infrastructure platforms already deploy for spam/abuse filtering.	RQ1: even a single exposure to AI comments converts 32.9% of readers; reducing exposure and the social-proof signal directly shrinks the conversion surface	Proposed; untested
L3	Campaign-level warning banner	When the comment section shows anomalously uniform directionality, display a section-level banner: “ <i>This comment section shows unusually uniform directionality and may include AI-coordinated activity.</i> ” In the absence of a detection trigger, retain a permanent low-key notice that “ <i>AI-generated comments may be present.</i> ”	RQ3: variance compression and OPPOSITE-reader pull are <i>section-level</i> phenomena that L1’s per-comment cue cannot communicate	Proposed; untested
L4	Diversity injection	When the top-ranked comments cluster on one direction, surface existing dissenting comments from elsewhere in the thread to the top, <i>reordering</i> rather than <i>generating</i> , to break the manufactured social-proof signal.	RQ3: MACM compresses variance by suppressing visible dissent; surfacing minority opinions is consistent with Asch (1956) ’s finding that conformity collapses with even one visible dissenter	Proposed; untested

Table 24: Mitigation Framework with empirical grounding. L1 is the only layer directly tested by the present RCT; L2–L4 are *motivated* by specific findings but not independently validated by this experiment.

but remain important for comprehensive platform policy.