How Do People Challenge Racial Stereotypes Online? Counter-Story Detection Across Reddit Communities

Warning: content in this paper may be upsetting or offensive.

Uma Sushmitha Gunturi[♠] **Jimin Mun**[♣] **Maarten Sap**[♣] **Maria Antoniak**[♡]

[♠]IBM [♣]Carnegie Mellon University [♡]University of Colorado Boulder

Abstract

Counter-storytelling is a powerful mechanism people use to challenge dominant narratives. Unlike other forms of counterspeech that have been widely studied in computational social science, counter-storytelling has largely been overlooked. Counter-stories are difficult to detect automatically; they are *relational* (defined with respect to expressions of racial stereotypes) and *structurally diverse* (drawing on stories that describe lived experiences, witnessed events, exemplars, and hypotheticals). We introduce a first framework for detecting and characterizing counter-storytelling against racial stereotypes at scale. This includes (1) a three-dimensional taxonomy grounded in narratology and Critical Race Theory and (2) a multi-stage pipeline that identifies relational pairs of stereotypes and counter-stories in noisy Reddit discourse. Using this pipeline, we annotate 25,549 Reddit posts across 615 communities and identify 1,312 counter-stories. Our analysis shows that speaker identity and post context shape how counter-stories are told. For example, in-group writers favor first-person testimony, often adopting the role of *self-reflective insiders*. Our work shows how computational methods can scale qualitative approaches to identify and characterize counter-storytelling as a contextual narrative practice, with implications for content moderation, narratology, and racial discourse analysis.[1]

1 Introduction

Counter-storytelling is the use of personal narratives to challenge dominant racial stereotypes—a practice that Critical Race Theory identifies as central to how marginalized communities resist racial oppression (Delgado and Stefancic, 2001; Solórzano and Yosso, 2002; Miles, 2019; Hunn et al., 2006). For example, in Fig. 1, an Asian user responds to a stereotypical post claiming *"all Asians look the same"*. They tell a counter-story that elegantly draws on lived experience, asserts in-group identity, and inverts the stereotype's premise. These stories can be more persuasive than factual arguments alone (Kubin et al., 2021; Broockman and Kalla, 2016; Lee and Lee, 2023) and can foster inclusive attitudes (Kalla and Broockman, 2020).

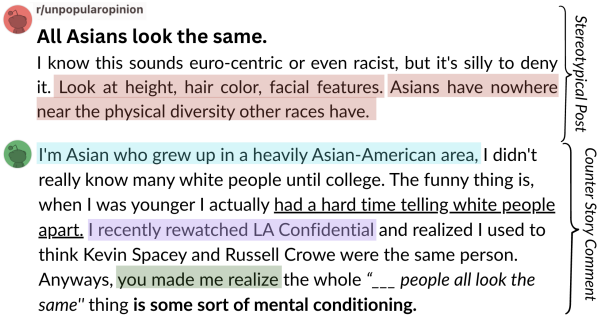

Figure 1: A motivating example of a counter-story responding to a stereotype-containing Reddit post (highlighted in red) on `r/unpopularopinion`. The example highlights the difficulty of identifying counter-storytelling: the narrator establishes a **narrative position** by recounting firsthand experience, asserts in-group **identity**, and directly addresses the original poster (**audience**). They also invert the racial premise of the stereotype and **reframe** the stereotypical perception as mental conditioning rather than a fact about the group.

Despite the power of these stories, most prior work in computational social science (CSS) has focused only on the broader discourse category of counter-speech, i.e., general responses against hate speech (Bonaldi et al., 2024; Fanton et al., 2021; Mun et al., 2023). But without a systematic understanding of counter-stories (who tells them, to whom, and from what position), we risk misinterpreting their social function. For example, moderation systems that cannot incorporate such context may flag counter-stories as toxic while failing to recognize the structural harms such stories contest

[1]Our pipeline code, race lexicon, stereotype carrier sentences, and annotation codebook are available at https://github.com/UmaGunturi/counter_story_detection

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 38622–38649
October 24-29, 2026 ©2026 Association for Computational Linguistics

(Fraser, 1990; Yang et al., 2025; Lee et al., 2024; Gunturi et al., 2024; Gligorić et al., 2024).

In this work, we computationally examine counter-storytelling as it unfolds at scale across hundreds of Reddit communities. We formalize *counter-story detection* as a *relational* narrative task in a multi-step pipeline (App. Fig. 6) that (1) identifies a stereotype-containing post, (2) identifies a story-containing response, and (3) determines whether the personal narrative challenges the stereotype.

Using this pipeline, we construct a large-scale dataset of counter-stories against racial stereotypes, comprising 25,549 Reddit posts (of which 3,600 are stereotypes) and 6,538 story-containing comments (of which 1,312 are counter-stories) across 615 unique subreddits, with each counter-story labeled along a three-dimensional taxonomy of *Narrative Position*, *Audience*, and *Identity* grounded in narratology and Critical Race Theory (§3). This dataset enables empirical investigation, at a scale prior resources could not support.

Using this dataset, we analyze the associations between the different narrative features, and we examine how community context and stereotype type are associated with counter-stories. We also perform a mixed-methods analysis to identify seven rhetorical roles, such as *outgroup allies*, *self-reflective insiders*, and *ingroup mentors*, by surfacing common feature combinations across our three taxonomy dimensions. Our findings demonstrate the potential of computational methods to advance understanding of narrative resistance online.

In summary, our contributions include:

- a three-dimensional taxonomy of Narrative Position, Audience, and Identity, grounded in narratology and Critical Race Theory (§3),
- a multi-stage pipeline that adapts and extends prior stereotype-detection and story-identification methods for noisy, large-scale online discourse, enabling counter-story annotation at a scale that human labeling cannot reach (§4),
- a large-scale annotated dataset of 3,600 stereotypical posts and 1,312 counter-story comments across 209 subreddits, validated by human annotators (§4),
- and a comprehensive analysis of online counter-storytelling patterns, with implications for content moderation and platform design (§5).

2 Related Work

Counter-stories in CRT and beyond. Counter-storytelling has been studied across many disciplines as a way marginalized communities contest dominant accounts of social life. These have included *critical race theory* (Delgado and Stefancic, 2001; Solórzano and Yosso, 2002; Ogbonnaya-Ogburu et al., 2020; Field et al., 2021), *sociology* (Shuman, 2006; Farrington, 2024), *literary theory* (Bal and van Boheemen, 2009; Chatman, 1978), and *public-sphere theory* on counterpublics (Fraser, 1990; Warner, 2002). Within CRT, Solórzano and Yosso (2002) describe "majoritarian stories"—dominant narratives that frame assumptions about marginalized groups as natural or self-evident. Race scholarship further illustrates how such dominant narratives are constructed, defining marginalized groups in relation to dominant groups and institutions (Omi and Winant, 2015; Le-Khac, 2020; Bonilla-Silva, 2014; Hanna et al., 2020). For example, scholarship has examined how racialized crime statistics constructed Blackness against a presumed-innocent white majority (Muhammad, 2010), how the "model minority" stereotype positioned Asian Americans in relation to Black civil rights claims (Wu, 2013), and how representations of Latino immigrants cast them as a demographic threat to the nation (Chavez, 2013). We draw on this scholarship (App. Table 9) to identify the recurring racial stereotypes through which these dominant narratives persist in online discourse today.

Counter-stories are a powerful mechanism to respond to such narratives by centering the lived experiences of people of color (Miles, 2019; Solórzano and Yosso, 2002; Hunn et al., 2006). Their function is therefore relational: it depends on the dominant narrative they contest, the social position from which they are told (Delgado and Stefancic, 2001), and the audience they address (Fraser, 1990; Warner, 2002). We use this same structure when constructing our annotation framework.

More broadly, prior empirical work has shown that personal narratives can shape how audiences interpret contested social issues, increasing respect across moral and political divides (Broockman and Kalla, 2016; Kalla and Broockman, 2020). Yet counter-storytelling has never been studied at scale. We offer the first computational account of how

counter-storytelling unfolds in online communities.

Counterspeech and counter narratives. Existing resources collect expert or crowd-authored counter-responses to hate speech, including CONAN (Chung et al., 2019), and related datasets spanning multiple languages and target communities (Chung et al., 2019; Fanton et al., 2021; Bonaldi et al., 2024; Yu et al., 2023). Counterspeech can include various rhetorical strategies, such as fact-correction, denouncing, humor, empathy, and perspective-taking (Chung et al., 2021b; Mun et al., 2023). These taxonomies capture argumentative and interpersonal strategies, but do not center counter-stories as a strategy or model narrative features.

Recent work has explored generating counter-narratives and evaluating them against expert-authored references (Qian et al., 2019; Tekiroğlu et al., 2020; Saha et al., 2022; Chung et al., 2021a; Furman et al., 2023; Mun et al., 2024). Rather than generating synthetic counter-stories, we examine how real people use counter-storytelling in their everyday lives online.

3 A Counter-Stories Framework

Counter-stories can vary widely in who tells them, how narrators position themselves in relation to the events they describe, and whom the story is addressed to. To characterize this variation systematically, we develop a three-dimensional taxonomy capturing *Narrative Position*, *Identity*, and *Audience* grounded in narratology and Critical Race Theory. Through iterative open coding of more than 300 counter-stories, we define preliminary dimensions (App. Fig. 4).

Narrative Position captures the narrator's relationship to the story being told: whether the narrator stands outside the story world, appears within it, or recounts their own experience as the protagonist of the story (Genette, 1980; Chatman, 1978; Fludernik, 2009).

Autodiegetic narrators are simultaneously narrator and protagonist. They recount firsthand experiences of discrimination or resistance in the first person, usually as members of the victim group.

Homodiegetic narrators are present in the story as witnesses rather than protagonists. This includes members of the victim group witnessing discrimination of others in their group, members of the dominant group recounting positive experiences with the victim group, and third-party observers witnessing interactions between groups.

Heterodiegetic narrators tell stories they were not present for, including: public exemplars, such as invoking public figures and historical events; second-hand narratives, which relay stories heard through trusted sources; and fictional narratives, which reference books, movies, and other media.

Metadiegetic narrators tell a story within a story, where the speaker's narrative includes another person's (e.g., quoted or reported) story.

Hypothetical narrators construct counterfactual scenarios, inviting the audience to reason through an alternative framing.

Audience captures the implied primary recipient of the counter-story, since the same story can function differently depending on whom it addresses (Fraser, 1990; Warner, 2002). Audience can include: direct rebuttal of the *Original Poster*; challenging or educating the *Dominant Group* who perpetuates the stereotype; offering solidarity, guidance, or critique to the *Victim Group*; addressing the broader *General/Mixed Audience*; and leaving the primary audience *Ambiguous*. Note that a single comment can address multiple audiences.

Identity captures the narrator's social perspective, following CRT's emphasis on positionality as constitutive of counter-storytelling (Solórzano and Yosso, 2002), who is speaking shapes what the story can claim and how it is likely to be received. Identity includes explicit self-disclosures of *race*, *gender*, and *connection* to the target group, separating narrators speaking as members of the stereotyped group from witnesses or allies.

4 Data and Annotation Pipeline

We apply our framework to a large dataset of Reddit posts and comments drawn from many diverse communities. We use a combination of lexicons, hand annotation, and LLM-assisted prediction to first identify posts containing stereotypes and then identify comments replying to those posts that contain counter-stories (App. Fig. 6 provides an overview of the full pipeline).

4.1 Data Collection and Filtering

Data source. We construct our dataset from two Reddit sources. For discussions prior to 2020, we use ConvoKit (Chang et al., 2020), which contains

Type	Label	#	%
Post	Stereotypes	3,600	14.1
	Asian	1,198	33.3
	Black	868	24.1
	Hispanic	71	2.0
	Other	1,463	40.6
	Non-stereotypes	21,949	85.9
Comment	Counter-stories	1,312	20.1
	Asian	484	36.9
	Black	351	26.8
	Hispanic	34	2.6
	Other	443	33.8
	Non-counter-stories	5,226	79.9

Table 1: Overview of the LLM annotated dataset.

∼948K subreddits. We also collect data via the Arctic Shift API[2] for conversations 2020–May 2026.

Race lexicon. We construct a lexicon of racial identity terms (e.g., "Black," "Asian"), slurs, and coded language for three focal groups: *Black*, *Asian*, and *Hispanic*. Lexicon construction and the full term list are in App. §A.

Community curation. We filter the ConvoKit corpus from ∼948K subreddits to a focused working set that is more likely to contain counter-storytelling. (1) Subreddits with fewer than 1,000 users, NSFW spaces, and quarantined communities are dropped. Within the remaining communities, we retain only those with at least one post containing a race-lexicon term. This leaves approximately 37,000 candidate subreddits spanning over 20 million posts. (2) We require each subreddit to contain at least 100 cumulative identity-keyword mentions, ensuring that race is a recurring rather than incidental topic. (3) For each subreddit, we compute a *match ratio*: the proportion of its posts (after the initial filtering step) that contain at least one lexicon term. We retain the top 1000 subreddits with match ratio ≥ 0.5. This step filters communities where race-related discussion is concentrated.

Post sampling and story detection. From these subreddits, we collect ∼150K lexicon-matched posts and retain only posts whose comment threads contain at least one story, applying StorySeeker (Antoniak et al., 2024), a classifier trained to detect stories on Reddit, to all top-level comments. We filter comments with <30 words and retain at most 500 story comments per post. This yields 90,061 posts with at least one story comment.

[2]https://github.com/ArthurHeitmann/arctic_shift

4.2 Identifying Posts with Racial Stereotypes

Concept-carrier semantic filter. Building on Mendelsohn and Budak (2025); Sap et al. (2020), we operationalize racial stereotypes as a set of *concepts* across three target groups: Black (e.g., *crime*), Asian (e.g., *model minority*), and Hispanic (e.g., *immigration*). We construct a concept inventory by defining 66 concepts across the three target groups. For each concept, we construct 10 carrier sentences (5 implicit and 5 explicit), which are written to express the stereotype with minimal semantic content (e.g., implicit: "They always seem to avoid work"; explicit: "They are lazy"). The full concept inventory, its sources from race scholarship, and example carriers per concept are in App. §D (Tables 9–13) and App. §E.3. For each concept, we compute centroid embeddings as the mean of its carriers (separately for its implicit and explicit variants) using Sentence-BERT (`all-MiniLM-L6-v2`; Reimers and Gurevych 2019). For each post, we extract the first sentence containing a race-lexicon keyword (matched at the token level) and compute its cosine similarity to every race-restricted concept centroid. We assign each post the concept with the highest score. Posts with $\max(\text{similarity}) > 0.30$ advance to the next stage.[3]

LLM assisted filtering. We label each of the 26K candidate posts (that passed the concept-carrier stereotype filter) with Claude Sonnet 4.6 using the few-shot prompt shown in App. §C.1. The model classifies whether each post expresses a racial stereotype, and if so, labels the victim group (Black, Asian, Hispanic, Other), stereotype type (e.g., Criminality), a certainty score (1–10) and a rationale. A total of 3.6K (14.1%) are labeled as expressing stereotypes (see Table 1 for detailed distribution stats). Note that the lexicon is used only to retrieve candidate posts via keyword matching; the target group and stereotype type are also assigned independently by the model from the full post content.

4.3 Labeling Counter-Stories

For the 3,600 stereotype-containing posts, we extract 6,538 story comments in total and prompt

[3]The 0.30 threshold was selected by inspecting samples at threshold values between 0.20 and 0.50; lower thresholds contained too many off-topic posts, higher thresholds excluded subtler stereotype posts.

Domain	#Subs	#CS	CS%	*Narrative Position*			*Audience*			*Identity*	
				Auto-diegetic	**Hetero-diegetic**	**Homo-diegetic**	**Original Poster**	**General/ Mixed**	**Victim Group**	**Target Group**	**Dominant Group**
Culture/Identity	66	430	20.9%	68%	8%	12%	98%	86%	35%	66%	5%
Advice/Education/Work	38	300	20.3%	72%	9%	10%	99%	80%	23%	33%	17%
Debate/Opinion	7	200	36.4%	34%	30%	14%	98%	98%	0%	10%	14%
General/Ask	19	140	24.0%	54%	15%	19%	100%	99%	3%	26%	31%
Stories	13	93	12.3%	65%	8%	12%	98%	80%	25%	25%	28%
Geography	28	53	22.2%	51%	15%	15%	98%	94%	6%	13%	13%
Humor	8	39	13.6%	38%	33%	18%	87%	100%	0%	3%	5%
Hobbies	14	24	16.6%	71%	17%	8%	96%	100%	0%	4%	4%
Misc	8	15	4.9%	73%	7%	13%	93%	87%	33%	33%	7%
Politics	5	12	13.8%	33%	42%	17%	92%	100%	0%	8%	0%
Beauty/Fashion	3	6	14.0%	67%	0%	0%	100%	100%	0%	33%	17%
Total	**209**	**1312**	**20.1%**	60%	14%	13%	98%	88%	19%	36%	14%

Table 2: Counter-stories grouped by community domain, ordered by counter-story count. **#Subs**: subreddits with ≥ 1 counter-story. **#CS**: total counter-stories. **CS%**: share of comments labeled counter-stories. Audience is multi-label, so columns may sum >100%. The remaining narrative positions (hypothetical, metadiegetic) are reported in App. Table 23. Speaker's gender and race breakdowns are reported separately in Table 16 and Table 21.

Annotator	Stereotype Agr.%	Stereotype κ	Counter-Story Agr.%	Counter-Story κ
Human	86	0.73	82	0.64
Sonnet 4.6	**88**	**0.76**	**86**	**0.72**

Table 3: Inter-annotator agreement (Fleiss' κ) across three annotators and model agreement with the human majority vote (Cohen's κ), on 50 stereotype posts and 100 counter-story comments.

Claude Sonnet 4.6 with the parent post, the comment, and our codebook (§3) to (i) decide whether the comment functions as a counter-story and (ii) annotate it along *Narrative Position*, *Audience*, and *Identity*. 1,312 comments (20.1%) are labeled as counter-stories, spanning 209 unique subreddits. We provide the full codebook in App. §B and the prompt in App. §C.2.

4.4 Human Validation

To assess how closely the model's labels align with human judgment, we conducted iterative human validation for both stereotype and counter-story labeling. Across multiple rounds, three human annotators independently annotated 50 stereotype examples and 100 counter-story examples, achieving substantial inter-annotator agreement ($\kappa = 0.73$ for stereotypes, $\kappa = 0.64$ for counter-stories) with disagreements used to refine the codebook and annotation guidelines. We then compared candidate LLM annotators against human labels; Claude Sonnet 4.6 showed the strongest agreement (Table 3 and App. Table 17). We further report precision/recall/F1 and find that the model's certainty scores are well calibrated with its accuracy (App. Tables 18, 19). A false-negative audit of ∼100 randomly sampled story comments found only ∼2% missed counter-stories, suggesting that the counter-stories missed by the pipeline are relatively uncommon (App. E.1).

5 Analysis

Given our annotated dataset of stereotype posts and counter-story comments spanning many diverse communities, we explore a series of questions that benefit from its scale.

5.1 Narrative Dimensions of Online Counter-stories

To understand the narrative diversity of counter-stories told online, we examine the distribution of the three narrative dimensions: narrative position, audience, and identity. We discuss these results as shown in Table 2 (App. Table 22 lists example subreddits per domain).

Counter-stories are most commonly told by auto-diegetic narrators. We observe that 60% of counter-stories found in our dataset were told using auto-diegetic narrative positions, followed by 14% heterodiegetic and 13% homodiegetic. The distribution differed based on community domains with Misc (73%), Advice/Education/Work (72%), Hobbies (71%), and Culture/Identity (68%) having the highest rate of autodiegetic narrators. Interest-

Stereotype Type	Asian	Black	Other
Cultural	74.3%	52.6%	76.5%
Appearance	10.6%	12.4%	5.7%
Criminality	2.9%	13.6%	6.6%
Dehumanization	4.7%	9.4%	3.9%
Other	3.4%	6.8%	2.9%
Intelligence	3.2%	3.1%	1.8%
Economic	0.9%	2.0%	2.6%
n posts	1,198	868	1,463

Table 4: Distribution of stereotype types by target group (column-wise within each group).

ingly, we observe higher heterodiegetic narrators than autodiegetic narrators in Politics, 42% and 33% respectively, and similar rates of hetero and autodiegetic narrators in Humor, 33% and 38%. The prevalence of autodiegetic narrators suggests that many counter-stories told online come from first-hand personal experiences. Moreover, the difference based on domain implies that narrative positions are contextual and community dependent. The elevated counter-story rate in Debate/Opinion (36.4%) is concentrated in *r/changemyview* (70.5%), a persuasion-oriented community (Antoniak et al., 2024) (App. Table 20).

Original posters are the most frequent audience. As shown in Table 2, the most frequent audience of counter-stories are the original posters who expressed the stereotype (98%), followed by a general/mixed audience (88%), and the victim group (19%). This pattern is statistically robust (McNemar's test, $p < 0.001$) for domains such as Culture/Identity, Advice/Education/Work, and Stories. Moreover, both original posters and general/mixed audience have high frequencies, suggesting that counter-storytellers often address both audiences in one story.

Counter-storytellers often belong to the target group. While most counter-storytellers do not disclose their group membership (53%), those who do are more often members of the target group (36%) than the dominant group (14%). Interestingly, we see a high percentage of in-group counter-storytelling in the Culture/Identity domain (66%), suggesting more in-group sharing and discussions. We hypothesize that in-group membership especially adds credibility to subjective experiences in this domain.

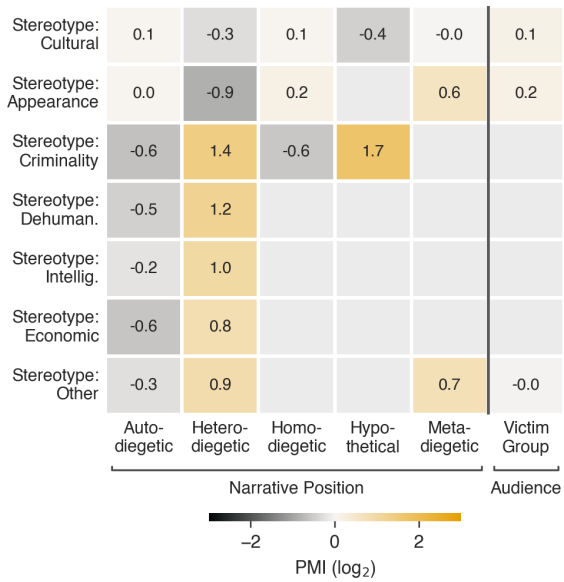

Figure 2: Pointwise mutual information (PMI, $\log_2$) between stereotype context and narrative dimensions across all the 1,312 counter-stories. Orange indicates feature pairs that co-occur more than chance; blank cells indicate no co-occurrence. Gray cells denote pairs with fewer than five co-occurrences ($n < 5$).

5.2 Influence of Context on Narrative Axes

To understand the relational aspects of counter-storytelling, we analyze the contexts in which counter-storytelling occurs. First, we observe the *post context*: the stereotype type and the target group of the stereotype. Second, we consider the *speaker context*: how the writer's disclosed relationship to the target group, race, and gender are associated with narrative dimensions. To surface these associations, we compute pointwise mutual information (PMI) scores between feature pairs across the 1,312 counter-stories (Figures 2 and 3).

Types of stereotype vary across racial identities. Stereotypes expressed in posts span seven categories across the three racial groups (see Table 4). Cultural stereotypes dominate across all groups, but notably, criminality and dehumanization stereotypes are most prevalent in Black-targeted posts (13.6%, 9.4%). By contrast, Asian-targeted posts are almost entirely cultural (74.3%) with comparatively little criminality (2.9%). See App. §E (Table 15) for stereotype distributions across all domains. We find that Hispanic-targeted content is sparsely represented in our data (71 stereotype posts and 34 counter-stories, ∼2% of each category). We therefore restrict cross-group interpretation to the Black and Asian groups and

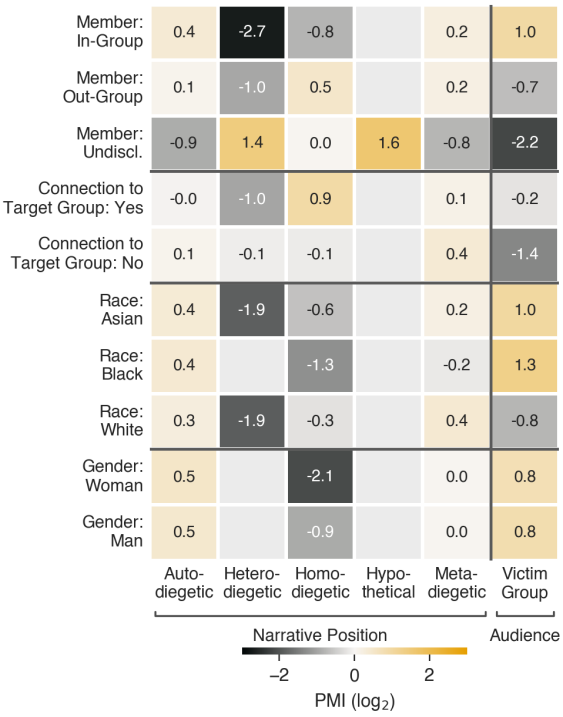

Figure 3: Pointwise mutual information (PMI, $\log_2$) between speaker context and narrative dimensions across all the 1,312 counter-stories. Orange indicates feature pairs that co-occur more than chance; blank cells indicate no co-occurrence. Gray cells denote pairs with fewer than five co-occurrences ($n < 5$). Additional feature pairs are reported in App. §E (Fig. 9).

report Hispanic statistics in App. §E.4.

Stereotypes are associated with narrative form. Cultural and appearance stereotypes sit near chance across narrative positions (PMI range -0.4 to 0.2, Fig. 2) (raw proportions in App. Table 14), suggesting they are countered in many different forms. Criminality and dehumanization stereotypes, however, co-occur more than chance with heterodiegetic narration (PMI 1.4, 1.2) and hypothetical narration (PMI 1.7, n=9). Counter-stories addressing criminality stereotypes frequently draw on public exemplars and documented cases (criminality $\times$ exemplar PMI 1.8; see App. Fig. 9). We also observe that the most stigmatizing stereotypes are less likely to be addressed *to* the targeted group: while cultural and appearance stereotypes sit near chance with victim-group address (PMI 0.1, 0.2), more stigmatizing stereotypes tend to orient toward the original poster or a general audience.

Speaker context is associated with narrative position. The writer's disclosed relationship to the stereotyped group is strongly associated with how they tell counter-stories (Fig. 3). Narrators in the target group co-occur with addressing the victim group (PMI 1.0). On the other hand, in-group members rarely co-occur with heterodiegetic (PMI -2.7) narration. This suggests that in-group members tend to speak *as* insiders *to* insiders, grounding their counter-story in lived experience and addressing those who share it.

Moreover, writers who report a personal connection without claiming membership tend to use homodiegetic witness narration (PMI 0.9) and who disclose no relationship to the targeted group show the opposite pattern: they co-occur more with heterodiegetic (PMI 1.4) and hypothetical narration (PMI 1.6), and less with victim-group address (PMI -2.2). This extends to narrative subtype: writers who disclose no membership lean on public exemplars (PMI 1.5) and fictional references (PMI 1.7, App. §E.2) (in-group $\times$ exemplar PMI -2.6; App. Fig. 9). Narrative-position distributions by writer identity and target group are shown in App. Fig. 8.

Similarly, among writers who disclose their race, Black and Asian writers are associated more with victim-group address (PMI 1.3, 1.0), unlike White writers (PMI -0.8) who are more likely to address the OP or a general audience.

Context alone does not predict counter-stories. We further examine whether *context* features and *speaker* features predict when a story-containing comment functions as a counter-story, by fitting logistic regression models over all story comments. Except in-group membership, no other feature emerges as a robust independent predictor, reinforcing that counter-storytelling may not be driven by any single factor but instead arise from *combinations* of speaker position, audience, stereotype type, and community context (App. E.5).

5.3 Rhetorical Roles in Counter-Storytelling

What kinds of rhetorical roles do counter-storytellers take on when challenging racial stereotypes? We build on prior work that clusters behavioral patterns of online community members (Yang et al., 2019; Sun et al., 2019; Mason and Ammari, 2026), and represent each of the 1,312 counter-stories as an 11-variable feature vector spanning the three taxonomy dimensions: narrative position, identity, and audience. We fit categorical Latent Class Analysis (LCA) (Collins and Lanza, 2013) models for $K = 2$ through $K = 12$ using the `stepmix` library (Morin et al., 2025) and

Role	n	%
Outgroup Ally	344	26.2
Self-Reflective Insider	276	21.0
Example Provider	221	16.8
Empathetic Reframer	207	15.8
Ingroup Mentor	173	13.2
Outgroup Reassurer	52	4.0
Context Provider	39	3.0
Total	**1,312**	**100.0**

Table 5: Distribution of counter-stories across the seven counter-storytelling roles identified by LCA. Total $N = 1{,}312$.

selected the model with the lowest BIC at $K = 8$. The first author then qualitatively interpreted the latent classes by reading the 20 most confidently assigned examples from each class and examining each class's label profile to understand the social and rhetorical role the commenter was taking on. Through an iterative process, we merged overlapping classes and assigned role names to the final classes. This yielded seven final counter-storytelling roles, shown in Table 5. The two merged classes include one almost purely heterodiegetic (Class 4, n=138) and one containing heterodiegetic and metadiegetic stories (Class 7, n=83), both citing external evidence. We qualitatively inspected these classes and merged them into one single class *"Example Provider"* (n= 221). The full BIC curve is in App. Fig. 7.

Outgroup Allies counter stereotypes from outside the targeted group by leveraging their dominant-group position to reject prejudice within their own communities ("*As a white person, I've heard my relatives make these kinds of comments, and they're simply wrong*"), often using personal experience to contest the stereotype ("*Some of the kindest, hardest-working people I know are immigrants...*").

Self-Reflective Insiders counter stereotypes from within the targeted group through introspective, often deeply personal stories about identity, family, appearance, relationships, and social expectations ("*I have felt the pressures of the stereotype of the studious, perfect Asian kid but it never really seemed to bother me. I knew deep down that I never really fit into that [...]*"). These comments are frequently autodiegetic as insiders narrate how stereotypes are absorbed, resisted, and reinterpreted from within the group.

Ingroup Mentors are members of the stereotyped group who respond to a fellow member's struggle by sharing their own parallel experience, then offering emotional support and concrete life advice. Rather than only refuting the stereotype, these counter-stories model survival, validation, and guidance for an in-group audience. Some examples include: shared experience ("*my dad was the same*"), reassurance ("*you deserve better*", "*I'm here*"), practical advice ("*speak to your doctor*", "*seek CBT*", "*my tip is...*"), well-wishing ("*I hope you're okay*").

Empathetic Reframers ask the original poster or broader audience to reinterpret a stereotyped situation with more care. They counter by shifting attention from racialized blame to alternative explanations, such as stress, grief, exhaustion, family dynamics, or situational constraint. For example, after an OP describes his Chinese partner and her family as "*more cold*," one commenter recalls, "*I had an emergency c-section and my immediate family came only to help me... With her being a new mom, sleep-deprived and breastfeeding, she's probably struggling. Give her some grace... she's probably just scared.*"

Outgroup Reassurers counter stereotypes from outside the targeted group by offering comfort and affirmation. They often speak directly to the stereotyped group and use their own experiences (e.g., dating, friendship, or family) to reassure stereotyped group. These comments are less about proving the stereotype wrong and more about reassuring the target that they are not alone. For example, one white commenter describes marrying an Indian partner despite family disapproval and tells the target: "*being Indian is not something to hate about yourself; find people who love you and do not make you feel ashamed.*"

Context Providers counter stereotypes by unpacking the broader conditions behind the behavior being racialized, such as labor exploitation, migration constraints, institutional failure, economic precarity, or unequal access to resources. For example, in response to the stereotype that "*Indians take jobs unfairly*," one commenter reframes: "*getting such jobs is not easy [...] my brother competed in an extremely selective admissions system, spent years studying long hours, and gave up much of regular school and social life to get there.*"

Example Providers counter stereotypes by offering concrete cases e.g., stories, jokes, film references, public examples (such as linking a Washington Post list of 100 racially miscast Hol-

lywood roles), or relational anecdotes as evidence. These comments challenge generalizations by supplying counterexamples in different forms to contest the stereotype.

6 Discussion

Counter-storytelling is a relational practice. Our findings confirm that counter-storytelling is fundamentally *relational*, consistent with CRT's emphasis on positionality (Solórzano and Yosso, 2002). Despite modeling a broad set of contextual and speaker-level features, only in-group membership remains a robust predictor of counter-storytelling after multiple-comparison corrections (§5.2). This pattern reinforces that whether a story functions as a counter-story depends on the interaction among speaker position, audience, and identity. Our latent role analysis further supports this interpretation, as the seven rhetorical roles emerge as recurring combinations of these dimensions. For example, *Outgroup Allies* combine out-group speaker position, heterodiegetic narration, and dominant-group address. At the same time, this relationality presents a challenge for computationally detecting counter-stories. Models that treat counter-storytelling as a property of isolated comments, without the surrounding stereotype post or interactional context, may mischaracterize how counter-stories function in practice (Gunturi et al., 2024; Hanna et al., 2020). Indeed, during our annotation process for counter-stories, we found that including the original post was essential for annotators to label the replies accurately. This also emphasizes the importance of community involvement and engagement, as interpretation of counter-stories relies on contextual knowledge. More broadly, computational approaches to racial discourse should therefore treat interaction and context as fundamental to understanding how racial narratives are reproduced and contested.

Implications for content moderation. Our findings highlight a challenge for moderation systems: counter-stories can resemble the harmful content they contest. The most common counter-stories in our dataset are first-person accounts from members of the targeted group, often addressed to the original poster and sometimes reusing the stereotype's language in order to invert or reject it. For example, a comment such as *"They called us 'lazy immigrants,' but my mother cleaned offices [...] so we could stay in school."* repeats the stereotype while turning it into a testimony of labor and sacrifice. Keyword- or toxicity-based systems that are neither relational nor identity-sensitive may flag such comments as harmful even when they function as resistance (Gligorić et al., 2024; Gunturi et al., 2024; Lee et al., 2024; Field et al., 2021). We find some confirmation of this concern through a toxicity audit of our dataset. Using Google's Perspective API with a 0.5 toxicity threshold, 9.4% of counter-stories were flagged as toxic, despite their role in responding to stereotypes, whereas 89.9% of stereotype-containing posts fell below the threshold. This problem arises because counter-stories derive their meaning from context. They are *relational*: a comment that appears to repeat a stereotype in isolation may be rejecting it in response to a parent post. They are also *positional*: the same narrative can carry different social meaning depending on the speaker's relationship to the targeted group (Zhou et al., 2023). Moderation pipelines that are neither relational nor identity-sensitive will consistently misclassify counter-stories as harmful, thereby suppressing important forms of community responses that challenge stereotypes (Fraser, 1990; Lee et al., 2024; Hanna et al., 2020).

Beyond moderation, our framework may also help surface authentic counter-stories so they are more likely to reach relevant audiences. Prior work has shown that a major barrier for counterspeakers is the perception that their efforts go unseen or have little impact (Mun et al., 2024; Ping et al., 2025). Increasing the visibility of counter-stories may therefore help sustain counterspeakers' engagement while promoting solidarity and cohesion within targeted communities.

7 Conclusion

Counter-storytelling is a critical tool for resisting racial stereotypes, yet its online practice remains poorly understood. We address this gap with a three-dimensional taxonomy grounded in narratology and critical race theory, capturing narrative position, audience, and identity. Using a multi-stage pipeline and LLM-assisted annotation, we construct a dataset of 25,549 Reddit posts across 615 communities, yielding 1,312 counter-story comments and 3,600 stereotypical posts. Our analysis reveals that counter-storytelling is diverse and relational, and we identify seven rhetorical roles that counter-storytellers commonly adopt.

8 Limitations

Our study has several limitations. First, our analysis focuses on English-language Reddit communities discussing racial stereotypes. Because our data is limited to English, the narrative forms, rhetorical conventions, and identity disclosures we observe may not generalize to counter-storytelling in other languages and cultural contexts, where norms around self-disclosure can differ. Our findings are also shaped by Reddit's threaded, long-form text format. Platform affordances may influence all three dimensions of our taxonomy: short-form video platforms such as TikTok may favor first-person testimony; quote-tweets and reply mechanisms on Twitter/X may shape whom a counter-story addresses; and private or semi-private forums may enable forms of in-group-directed storytelling that are less visible in public data. Counter-storytelling may likewise take different forms on other platforms with different interactional norms (e.g., Twitter, TikTok, private forums), in other linguistic and cultural contexts, or when responding to stereotypes about other marginalized identities such as gender, disability, religion, or sexuality. Extending our framework across languages, platforms, and identity axes is therefore an important direction for future work.

Second, our dataset construction reflects methodological trade-offs. Our stereotype discovery pipeline combines lexicon-based retrieval, LLM-assisted labeling, and concept–carrier sentence expansion to identify posts containing stereotypes at scale. While this improves coverage beyond simple keyword matching, it may still miss subtler or community-specific stereotypes that do not align with our seed concepts, while LLM labeling may introduce false positives or biases in what is recognized as stereotypical content.

Third, identity annotations rely only on explicit self-disclosure in text. Our analysis therefore captures how users position their identities within the conversation, rather than the full demographic composition of participants.

9 Ethical Considerations

The data we examine in this paper is public but sensitive. This tension leads to a series of tradeoffs between scientific transparency and reproducibility, privacy and agency of the Reddit users, and safety of our annotators and readers. Many of the stories and counter-stories in this study contain very personal information about the user including aspects of their identity. We determined that review for this study was not required by our institution's Internal Review Board (IRB), as users do not have an expectation of privacy, but the nature of this topic and data requires careful consideration.

Data decisions. Both stereotype-containing posts and storytelling comments may contain distinctive personal details that could enable re-identification. To ensure counter-storytellers' privacy and agency, we therefore do not release raw text, user identifiers, or post-comment pairs. However, to support research transparency while minimizing redistribution of sensitive narratives, we will make the annotated dataset available upon request for research purposes only. All requests will be reviewed with attention to the goals of the research, the intended use of the data, and safeguards for handling sensitive personal narratives. We publicly release the annotation codebook, pipeline code, race lexicon, and carrier sentences needed to support reproducibility.

All examples used in the paper were de-identified; no usernames or other identifying information are included. Importantly, we never try to *predict* identity features of the users and only use explicit disclosures ("*As an Asian woman, ...*") to categorize the texts; these public disclosures were intentional choices of the users and are significant to building our understanding of counter-storytelling. We also minimize our data collection, and we do not collect additional data about these users; e.g., we do not track individual users across conversations and communities, and we do not attempt to construct profiles of individual users.

Risks. We note two potential misuses: (1) the codebook could be mis-used to craft (manually or automatically) deceptive narratives, and (2) automated detection of counter-storytelling could be misused by bad actors to locate, monitor, or suppress resistance by marginalized communities.

Safety. Our annotators (including the authors) and to a lesser extent, our readers, are exposed to extremely explicit and toxic racist stereotypes through the data. Annotators discussed taking breaks and prioritizing their mental health. Unfortunately, the nature of the research questions posed in this work necessitate reading and contending with many different kinds of racial stereotypes, which often co-occur with misogynist and other upsetting language. We provide a warning at the

beginning of the paper, and we minimize the toxicity of the examples shown in the paper, but given our research focus, we cannot avoid directly engaging with this content. We also acknowledge our positionality in this work: some authors belong to the stereotyped communities represented in our study, which informed our engagement with and interpretation of these narratives, while some authors do not belong to these groups.

Use of AI Assistance We used AI assistants for limited support during coding, analysis, and writing. All data decisions, annotations, citations, statistical analyses, interpretations, and final writing decisions were reviewed and made by the authors.

Acknowledgments

We are especially grateful to Dr. Long Le Khac for the many generous discussions that helped shape and develop our annotation framework. We thank Teagan Johnson, Advait Deshmukh, and colleagues at Carnegie Mellon University and the University of Colorado Boulder for helpful feedback throughout the project. We also thank the anonymous reviewers for their constructive comments.

References

Juan José Alonzo. 2009. *Badmen, Bandits, and Folk Heroes: The Ambivalence of Mexican American Identity in Literature and Film*. University of Arizona Press.

Maria Antoniak, Joel Mire, Maarten Sap, Elliott Ash, and Andrew Piper. 2024. Where do people tell stories online? story detection across online communities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7104–7130, Bangkok, Thailand. Association for Computational Linguistics.

M. Bal and C. van Boheemen. 2009. *Narratology: Introduction to the Theory of Narrative*. Narratology: Introduction to the Theory of Narrative. University of Toronto Press.

Lee Bebout. 2016. 1. what did they call them after they called them "greasers"?: A genealogy and taxonomy of the mexican other. In *Whiteness on the Border*, pages 33–72. New York University Press.

Donald Bogle. 2001. *Toms, coons, mulattoes, mammies, and bucks: An interpretive history of Blacks in American films*. Bloomsbury Publishing.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for counterspeech against hate: A survey and how-to guide. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, Mexico City, Mexico. Association for Computational Linguistics.

Eduardo Bonilla-Silva. 2014. *Racism without racists color-blind racism and the persistence of racial inequality in America*, 4th ed. edition. Rowman & Littlefield Pub., Lanham, Md.

David E. Broockman and Joshua L. Kalla. 2016. Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, 352:220 – 224.

Mrinalini Chakravorty. 2014. *In stereotype: South Asia in the global literary imaginary*. Columbia University Press.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Seymour Benjamin Chatman. 1978. *Story and discourse: Narrative structure in fiction and film*. Cornell university press.

L. Chavez. 2013. *The Latino Threat: Constructing Immigrants, Citizens, and the Nation, Second Edition*. Stanford University Press.

Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021b. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Linda M Collins and Stephanie T Lanza. 2013. *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. John Wiley & Sons.

Richard Delgado and Jean Stefancic. 2001. *Critical Race Theory: An Introduction*. NYU Press.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Re'Nyqua Farrington. 2024. Twitter as a counter-storytelling site for students of color working to abolish the police. *Equality, Diversity and Inclusion: An International Journal*, 43(3):551–567.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Monika Fludernik. 2009. *An introduction to narratology*. Routledge.

Nancy Fraser. 1990. Rethinking the public sphere: A contribution to the critique of actually existing democracy. *Social Text*, (25/26):56–80.

Damián Ariel Furman, Pablo Torres, José Raúl Rodríguez Rodríguez, Diego Letzen, María Vanina Martínez, and Laura Alonso Alemany. 2023. An initial exploration of how argumentative information impacts automatic generation of counter-narratives against hate speech. In *Arg&App 2023: International Workshop on Argumentation and Applications*.

Racquel J. Gates. 2018. *Double Negative: The Black Image and Popular Culture*. Duke University Press.

G. Genette. 1980. *Narrative Discourse: An Essay in Method*. Cornell paperbacks. Cornell University Press.

Kristina Gligorić, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959, Mexico City, Mexico. Association for Computational Linguistics.

Uma Sushmitha Gunturi, Anisha Kumar, Xiaohan Ding, and Eugenia H. Rho. 2024. Linguistically differentiating acts and recalls of racial microaggressions on social media. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Alex Hanna, Remi Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 501–512, New York, NY, USA. Association for Computing Machinery.

Melissa V Harris-Perry. 2011. *Sister citizen: Shame, stereotypes, and Black women in America*. Yale University Press.

Lisa R. Merriweather Hunn, Talmadge C. Guy, and Elaine Mangliitz. 2006. Who can speak for whom? using counter-storytelling to challenge racial hegemony. In *Proceedings of the Adult Education Research Conference*. New Prairie Press.

Joshua L. Kalla and David E. Broockman. 2020. Reducing exclusionary attitudes through interpersonal conversation: Evidence from three field experiments. *American Political Science Review*, 114:410 – 425.

Emily Kubin, Curtis Puryear, Chelsea Schein, and Kurt Gray. 2021. Personal experiences bridge moral and political divides better than facts. *Proceedings of the National Academy of Sciences*, 118(6):e2008389118.

Long Le-Khac. 2020. *Giving Form to an Asian and Latinx America*. Stanford University Press.

Cinoo Lee, Kristina Gligorić, Pratyusha Ria Kalluri, Maggie Harrington, Esin Durmus, Kiara L. Sanchez, Nay San, Danny Tse, Xuan Zhao, MarYam G. Hamedani, Hazel Rose Markus, Dan Jurafsky, and Jennifer L. Eberhardt. 2024. People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise. *Proceedings of the National Academy of Sciences*, 121(38):e2322764121.

Jeehyun Jenny Lee and Jin Lee. 2023. #stopasianhate on tiktok: Asian/american women's space-making for spearheading counter-narratives and forming an ad hoc asian community. *Social Media + Society*, 9(1):20563051231157598.

Robert G. Lee. 1999. *Orientals: Asian Americans in Popular Culture*. Temple University Press.

José E Limón. 1973. Stereotyping and chicano resistance: An historical dimension. *Aztlán: A Journal of Chicano Studies*, 4(2):257–270.

Hazel Markus and Paula M. L. Moya. 2010. *Doing race : 21 essays for the 21st century*. W. W. Norton & Co.

Sherry Mason and Tawfiq Ammari. 2026. Racism, resistance, and reddit: how popular culture sparks online reckonings. *Information, Communication & Society*, 0(0):1–18.

Julia Mendelsohn and Ceren Budak. 2025. When people are floods: Analyzing dehumanizing metaphors in immigration discourse with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8079–8103, Vienna, Austria. Association for Computational Linguistics.

James Miles. 2019. Historical silences and the enduring power of counter storytelling. *Curriculum Inquiry*, 49(3):253–259.

Sacha Morin, Robin Legault, Félix Laliberté, Zsuzsa Bakk, Charles-Édouard Giguère, Roxane de la Sablonnière, and Éric Lacourse. 2025. Stepmix: A python package for pseudo-likelihood estimation of generalized mixture models with external variables. *Journal of Statistical Software*, 113(8):1–39.

Khalil Gibran Muhammad. 2010. *The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America*. Harvard University Press.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777, Singapore. Association for Computational Linguistics.

Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' perspectives: Unveiling barriers and ai needs in the fight against online hate. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical race theory for hci. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–16, New York, NY, USA. Association for Computing Machinery.

Michael Omi and Howard Winant. 2015. *Racial formation in the United States*, third edition. edition. Routledge/Taylor & Francis Group, New York.

Kaike Ping, Anisha Kumar, Xiaohan Ding, and Eugenia H. Rho. 2025. Behind the counter: Exploring motivations and barriers of online counterspeech writing. *ACM Trans. Comput.-Hum. Interact.*, 32(5).

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5157–5163. International Joint Conferences on Artificial Intelligence Organization.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Amy Shuman. 2006. *Storytelling rights: The uses of oral and written texts by urban adolescents*. 11. Cambridge University Press.

Daniel G. Solórzano and Tara J. Yosso. 2002. Critical race methodology: Counter-storytelling as an analytical framework for education research. *Qualitative Inquiry*, 8(1):23–44.

Lu Sun, Robert E. Kraut, and Diyi Yang. 2019. Multilevel modeling of social roles in online micro-lending platforms. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Yuval Taylor and Jake Austen. 2012. *Darkest America: Black minstrelsy from slavery to hip-hop*. WW Norton & Company.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Linda G Tucker. 2009. *Lockstep and dance: Images of Black men in popular culture*. Univ. Press of Mississippi.

Michael Warner. 2002. *Publics and Counterpublics*. Zone Books.

Ellen D. Wu. 2013. *The Color of Success: Asian Americans and the Origins of the Model Minority*. Princeton University Press, Princeton.

Diyi Yang, Robert E. Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–14, New York, NY, USA. Association for Computing Machinery.

Yukun Yang, Robin Lange, Yidi Zhang, and Joseph B Walther. 2025. How social responses to online hate messages affect hatefulness. *New Media & Society*, 0(0):14614448251396951.

Xinchen Yu, Ashley Zhao, Eduardo Blanco, and Lingzi Hong. 2023. A fine-grained taxonomy of replies to hate speech. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7275–7289, Singapore. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

A Race Lexicon

Our race lexicon comprises identity keywords (neutral terms for self-identification (e.g., "Black," "Asian American," "Hispanic"), slurs (sourced from `https://slurs.info/slurs/`), and dog-whistles (sourced from `https://dogwhistles.allen.ai/`) for three racial groups. Tables 6–8 present example terms from each racial group; the complete lexicon is available in our supplementary materials.

Category	Example Terms
Identity keywords	black, african american, bipoc, african, african descent, afro-american
Dog-whistles	jogger(s), ghetto(s), urban, welfare queen(s), thug(s), black-on-black crime, absent fathers, 13%, 13/50
Slurs	[We include 600+ slurs in our lexicon but omit them here due to their highly offensive nature. Full list available upon request for research purposes.]

Table 6: Black/African race lexicon. Identity keywords: $n = 9$; Dog-whistles: $n = 28$; Slurs: $n = 616$.

B Counter-Story Annotation Codebook

Fig. 5 displays the codebook used to annotate counter-narratives along three dimensions of our taxonomy (Fig. 4): narrative position, audience, and author identity.

C LLM Annotation Prompts

C.1 Stereotype Detection

We show the few-shot prompt below; the zero-shot variant is identical with the **Examples** block removed.

Category	Example Terms
Identity keywords	latino, latina, latinx, hispanic, mexican american, chicano, puerto rican, cuban, central american, south american
Dog-whistles	secure the border, illegal aliens, criminal aliens, anchor baby/babies, stealing jobs, cartel, deport them all, press 2 for spanish, mexifornia
Slurs	[We include 170+ slurs in our lexicon but omit them here due to their highly offensive nature. Full list available upon request for research purposes.]

Table 7: Hispanic race lexicon. Identity keywords: $n = 15$; Dog-whistles: $n = 24$; Slurs: $n = 175$.

Category	Example Terms
Identity keywords	asian, asian American, east asian, south asian, southeast asian, indian, chinese, japanese, korean, vietnamese, filipino, pakistani, bangladeshi, indonesian
Dog-whistles	model minority, china virus, kung flu, wuhan virus, foreigner, tiger mom, abolish birthright citizenship, ccp influence, academic cheats, tech job thieves, high iq, stem overlords
Slurs	[We include 340+ slurs in our lexicon but omit them here due to their highly offensive nature. Full list available upon request for research purposes.]

Table 8: Asian race lexicon. Identity keywords: $n = 19$; Dog-whistles: $n = 25$; Slurs: $n = 345$.

System

You are an expert content analyst for racial stereotypes in online discourse. You are analyzing Reddit posts pre-filtered for racial/ethnic content. Your job is to determine whether the post expresses a racial stereotype.

Definition. A racial stereotype is a claim attributed to a racial/ethnic group that is either: (a) a generalization applied broadly to the group, OR
(b) an individual anecdote framed as representative of the group, OR
(c) humor, jokes, or slurs that dehumanize or demean a racial group.
Stereotypes can be negative, positive ("model minority"), or use coded language ("those neighborhoods").

Key Rule. The author must ENDORSE or ASSERT the stereotype. Reporting, quoting, or refuting someone else's stereotype does NOT count — even if the post is about stereotypes.

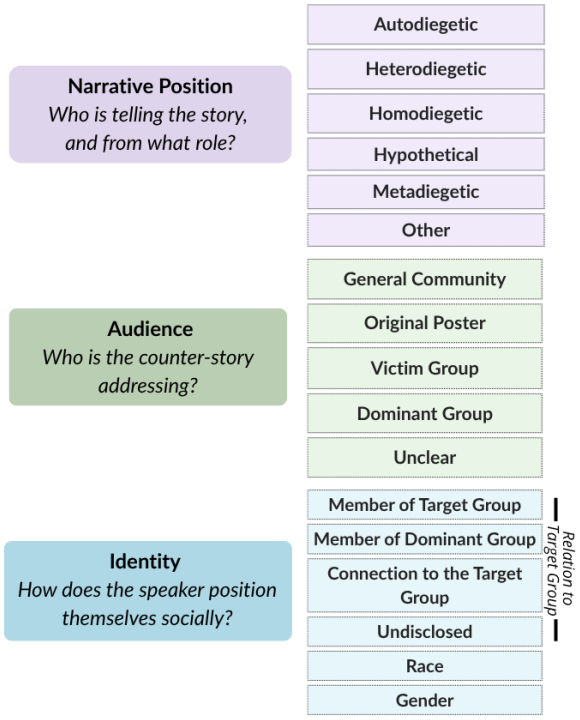

Figure 4: Our three-dimensional taxonomy for characterizing counter-stories: **Narrative Position** captures how the narrator formally relates to the story, **Audience** captures who the counter-story addresses; comments may address multiple audiences, and **Identity** captures how the speaker positions themselves relative to the stereotyped group, including self-disclosed demographic attributes.

Classify YES when:
- The author generalizes about a racial/ethnic group, explicitly or implicitly
- Individual/family anecdotes are framed as "typical" or representative of the group
- Positive-framed stereotypes (model minority, "naturally athletic")
- Coded/dog-whistle language when the racial group is clear from context
- In-group authors count — judge the content, not the author's identity
- Hedged claims count if the underlying assertion is endorsed
- Jokes whose punchline relies on racial slurs or dehumanizes a racial group count even without explicit trait attribution

Classify NO when:
- Author is reporting, quoting, analyzing, or refuting a stereotype
- Post cites statistics/data without claiming inherent group traits
- Post describes a specific individual without framing them as representative
- Fictional/hypothetical contexts without stereotypical claims
- Sarcastic reversals clearly mocking the stereotype

Examples.
Example 1 — YES: "Why are Black people always so loud at restaurants lol"
`{"stereotype":"Yes", "target_group":"Black", "stereotype_type":"cultural", "rationale":"Explicit generalization.", "certainty":9}`
Example 2 — NO: "My professor talked about how Asian students are stereotyped as model minorities, it was an eye-opening lecture about implicit bias."
`{"stereotype":"No", "target_group":"None", "stereotype_type":"none", "rationale":"Reports critically; author does not endorse.", "certainty":9}`
Example 3 — YES: "Every Black guy I've met plays basketball, it's just a fact lol"
`{"stereotype":"Yes", "target_group":"Black", "stereotype_type":"cultural", "rationale":"Anecdote asserted as general fact.", "certainty":9}`
Example 4 — NO: "My dad keeps saying all Mexicans are criminals but he's never met one. Tired of this ignorance."
`{"stereotype":"No", "target_group":"None", "stereotype_type":"none", "rationale":"Quotes and refutes.", "certainty":9}`
Example 5 — YES: "I don't want to be racist but there's a reason property values drop when Black people move in..."
`{"stereotype":"Yes", "target_group":"Black", "stereotype_type":"economic", "rationale":"Coded reference; hedging confirms awareness.", "certainty":8}`
Example 6 — NO: "Oh yeah, all Asians are good at math and never have feelings outside of studying, right?"
`{"stereotype":"No", "target_group":"None", "stereotype_type":"none", "rationale":"Sarcastic reversal mocking the stereotype.", "certainty":8}`

Certainty Scale.
- 9–10: Unambiguous
- 7–8: Clear but requires interpretation
- 5–6: Ambiguous about endorsement or scope
- 1–4: Not a stereotype

Analysis Steps.
1. Target group
2. Trait/claim attributed — if none, check for slurs or dehumanization
3. Endorsement: asserting vs. reporting?
4. Verdict

Output ONLY valid JSON:
```
{"reasoning":"<2-3 sentences>",
"stereotype":"Yes/No",
"target_group":
"<Black/Asian/Mexican/White/Other/None>",
"stereotype_type":
"<dehumanization/criminality/intelligence/
appearance/cultural/economic/other/none>",
"rationale":"<1-2 sentences>",
"certainty":<1-10>}
```

1. NARRATIVE POSITION

Captures the narrator's relationship to the story being told—whether the narrator stands outside the story world, appears within it, or recounts their own experience as its protagonist (Genette, 1980; Chatman, 1978).

1. **Autodiegetic.**
   Definition: The narrator is simultaneously narrator and protagonist, recounting firsthand experiences of discrimination or resistance in the first person, most often as a member of the stereotyped group.
   Example: "As a Black diner, the waiter checked on the white couple beside us five times and never once on us."

2. **Homodiegetic.**
   Definition: The narrator is present in the story as a witness rather than its protagonist. It could be:

   - a victim-group member witnessing discrimination toward others in their group,
   - a dominant-group member recounting positive experiences with the victim group,
   - a third-party observer of interactions between groups.

   Example: "I waited tables for years and watched servers breeze past the Black families in their section."

3. **Heterodiegetic.**
   Definition: The narrator recounts stories they were not present for, spanning three subtypes:

   - *Public exemplars*—public figures, historical events, or well-known cases.
   - *Second-hand*—stories heard through trusted sources.
   - *Fictional*—books, films, or other media invoked to challenge the stereotype.

   Examples:

   - (*exemplar*) *"Condoleezza Rice writes about a principal thanking her 'for showing us that not all black people are stupid'."*
   - (*second-hand*) *"My coworker told me her landlord refused to rent to her the second he heard her accent over the phone."*
   - (*fictional*) *"Read Native Speaker; it captures growing up Asian American better than any argument."*

4. **Metadiegetic.**
   Definition: The narrator tells a story within a story, embedding another person's (e.g., quoted or reported) account.
   Example: "My grandmother used to tell us about a waiter who said 'we don't serve your kind here'...she'd describe how the whole room just went quiet."

5. **Hypothetical.**
   Definition: The narrator constructs a counterfactual scenario, inviting the audience to reason through an alternative framing.
   Example: "Imagine a white teenager getting stopped and frisked walking home...would anyone call that 'routine'?"

2. AUDIENCE

Captures the implied primary recipient of the counter-story, since the same story can function differently depending on whom it addresses (Fraser, 1990; Warner, 2002). A single comment can address multiple audiences.

- **Original Poster (Direct Rebuttal).** Directly responds to the OP's stereotyping statement through personal confrontation.
- **Dominant Group.** Challenges or educates those who perpetuate the stereotype.
- **Victim Group (In-group).** Offers solidarity, guidance, or critique to in-group members.
- **General/Mixed Audience.** Speaks to the broader community; educational or contextual.
- **Ambiguous/Multiple.** The primary audience is unclear or addressed simultaneously.

3. AUTHOR IDENTITY

Captures the narrator's social perspective. Identity comprises explicit self-disclosures, separating narrators who speak from lived membership in the stereotyped group from those who speak as witnesses or allies.

- **Race:** White, Asian, Black, Hispanic, Other, Unknown.
- **Gender:** Man, Woman, Non-Binary, Trans, Unknown.
- **Member of Target Group:** Yes / No / Unknown.
- **Member of Dominant Group:** Yes / No / Unknown.
- **Personal Connection to Target Group:** Yes / No / Unknown (only if not a member of the target group).

COMMENTS THAT ARE *NOT* COUNTER-NARRATIVES

A counter-narrative must actively challenge the specific stereotype in the post. A comment is *not* a counter-narrative if it:

- agrees with, amplifies, or reinforces the stereotype (even with surface-level disagreement);
- is a pseudo-counter-narrative that appears to challenge the stereotype but ultimately reinforces it;
- changes the subject entirely or makes unrelated observations;
- offers general commentary about racism without specifically countering the post's stereotype; or
- contains no narrative (pure argument, data citation, or abstract meta-commentary).

Figure 5: Codebook for counter-story annotation.

User

Subreddit: `r/{subreddit}`

Post Title: "{title}"

Original Post: "{post_text}"

C.2 Counter-Story Annotation

For counter-story annotation, we experimented with both zero-shot and few-shot prompting and observed no significant difference in agreement with the human labels; we therefore use the codebook-guided prompt below.

System

You are an expert annotator for counter-narrative research on racial stereotypes in online discourse.

What is a Counter-Narrative? A counter-narrative is a story or response that:
- Opposes or subverts a dominant/majoritarian narrative (here: a racial stereotype)
- Centers the lived experience or dignity of the targeted group
- Reframes the representation in a way that undermines the stereotype's legitimacy

Counter-narratives oppose hate/stereotype content through strategies like presenting contradicting facts or lived experience, pointing out hypocrisy or logical contradiction, humanizing the targeted group, or reframing the narrative from the targeted group's perspective.

Key rule. A counter-narrative must actively challenge the specific stereotype in the post, not merely agree with or amplify the stereotype, change the subject entirely, make unrelated observations, or add general commentary about racism without specifically countering the stereotype. Comments that agree with or reinforce the stereotype, or pseudo-counter-narratives that appear to challenge but ultimately reinforce the stereotype, are NOT counter-narratives.

Annotation Dimensions.

0a. CONTAINS NARRATIVE?
- Does the comment include a story, personal experience, observation, or retelling of events?
- Note that pure arguments, data citations, or abstract meta-commentary WITHOUT story elements = No.
- Note that a comment can contain multiple narratives on different topics—that still counts as Yes.

Yes / No

0b. IS COUNTER-NARRATIVE? Does this comment challenge, contradict, humanize, or provide an alternative perspective to the racial stereotype in the post? Mark No if it agrees with/reinforces the stereotype, does not meaningfully challenge and counter it, doesn't contain a narrative, or makes general anti-racism commentary without addressing the post's stereotype. *Yes / No*

1. NARRATIVE POSITION (if 0a = No, use "N/A"):
- *Autodiegetic*—narrator is protagonist with direct personal experience
- *Homodiegetic*—narrator present as witness/observer, not protagonist
- *Heterodiegetic*—third-person stories narrator was NOT present for
- *Metadiegetic*—story within a story, nested/multiple narrative levels
- *Hypothetical*—hypothetical scenarios or thought experiments
- *Other*—narrative position present but doesn't fit above
- *N/A*—no narrative position

2. NARRATIVE SUBTYPE (only if Heterodiegetic, else "N/A"; comma-separated if multiple):
- *Second-hand*—stories from people narrator knows
- *Exemplar/Celebrity*—stores about historical figures, celebrities, public cases
- *Fictional*—stories related to fiction, media, books, movies

3. AUDIENCE (all that apply, comma-separated):
- *Dominant Group* (challenging/educating those who perpetuate stereotypes)
- *Victim Group* (solidarity, guidance, or critique to in-group)
- *Original Poster* (directly responding to OP)
- *General/Mixed* (broader Reddit community)
- *Unclear*

4. IDENTITY – RACE (stated or clearly implied): White / Black / Asian / Hispanic / Other / Unknown

5. IDENTITY – GENDER (stated or clearly implied): Man / Woman / Non-Binary / Trans / Unknown

6. MEMBER OF TARGET GROUP? Is the narrator a member of the stereotyped group? Yes / No / Unknown

7. MEMBER OF DOMINANT GROUP? Is the narrator a member of the dominant (typically white/-majority) group? Yes / No / Unknown

8. CONNECTION TO TARGET GROUP (only if NOT a member of target group):
Is the narrator a personal connection (family, partner, close friend) to the target group? Yes / No / Unknown / N/A

9. RATIONALE (step-by-step reasoning—do this FIRST, before fixing your verdicts):
(a) What is the specific racial stereotype in the post?
(b) What is the commenter actually saying/doing? Summarize the comment's stance.
(c) Is there a narrative (personal experience, retelling, observation)? What kind?
(d) Does the comment meaningfully challenge the stereotype in (a)? How, or why not?
(e) Any edge cases, ambiguity, sarcasm, or pseudo-counter-narrative features?
Be concrete, reference actual content from the comment.

10. CERTAINTY (1–10), for the `is_counter_narrative` verdict: 1–3 = very unsure; 4–6 = moderate; 7–9 = high; 10 = unambiguous

Output ONLY a valid JSON object. Fill `rationale` FIRST (chain-of-thought), then use that reasoning to fill the verdicts:

```
{"rationale":"...",
"contains_narrative":"Yes/No",
"is_counter_narrative":"Yes/No",
```

```
"narrative_position":"...",
"narrative_subtype":"...",
"audience":"...",
"identity_race":"...",
"identity_gender":"...",
"member_target_group":"...",
"member_dominant_group":"...",
"connection_target_group":"...",
"certainty":1-10}
```

User

Subreddit: `r/{subreddit}`
Post Title: "{title}"
Original Post (contains racial stereotype about `{target_group}`): "{post_text}"
Comment to analyze: "{comment_text}"

D Stereotype Concepts and Carrier Sentences

We follow the methodological approach outlined in (Mendelsohn and Budak, 2025) for generating carrier sentences associated with stereotype *concepts*. Similar to their framework, we compile source stereotype sentences for each concept from literature (see Table 9 for more details) per race group and use them as the foundation for generating new implicit and explicit carrier sentences. To generate these sentences, we provided GPT–4o with the standardized prompt shown below.

System

A **concept** is the stereotype category or stereotypical association to be expressed. A **carrier sentence** is a short, generic sentence that expresses that concept with minimal additional semantic detail. Given a concept and the example stereotypes, generate 5 implicit and 5 explicit versions of the stereotypical carrier sentences per concept using the example sentences provided. Limit it to 5 and not more than that. Do not generate any other information or reasoning.

User

Concept: `{concept}`
Example sentences: `{source_sentences}`

Group	Sources
Black	(Muhammad, 2010); (Harris-Perry, 2011); (Gates, 2018); (Taylor and Austen, 2012); (Tucker, 2009); (Bogle, 2001)
Asian	(Wu, 2013); (Lee, 1999); (Markus and Moya, 2010); (Chakravorty, 2014)
Hispanic	(Chavez, 2013); (Bebout, 2016); (Alonzo, 2009); (Limón, 1973)

Table 9: Sources from which we drew dominant-stereotype concepts. Concepts and carrier sentences were constructed from stereotype expressions documented in these works.

E Additional Information about the Dataset

E.1 False-Negative Audit for Missed Counter-Stories

We conducted a small-scale false-negative audit focusing on two likely sources of missed counter-stories: (i) story-comments under stereotype posts that were labeled as non-counter-stories by the LLM, and (ii) story-comments under posts scoring just below the stereotype retrieval threshold (0.30). We manually inspected approximately 100 comments from these audit pools and found only $\sim$2% missed counter-stories, providing some evidence that missed counter-stories are relatively uncommon. The missed cases tended to challenge stereotypes indirectly. For example, one countered an *"Asian students cheat"* stereotype by recounting widespread cheating at the commenter's own school (rather than explicitly addressing it), reframing the behavior as situational rather than racial; Sonnet correctly classified this comment as a first-person narrative but failed to recognize that the story challenges a stereotype. Comprehensive false-negative estimation remains challenging because counter-stories are inherently sparse and require jointly identifying a stereotype, a story, and a countering response.

E.2 Fiction and Media as Counter-Storytelling Resources

While most counter-stories in our dataset draw on lived experience, we also observed a small number of cases in which writers referenced fictional, literary, or media texts when responding to stereotypes. We do not analyze this phenomenon systematically, but include these examples to document an additional form of counter-storytelling present in

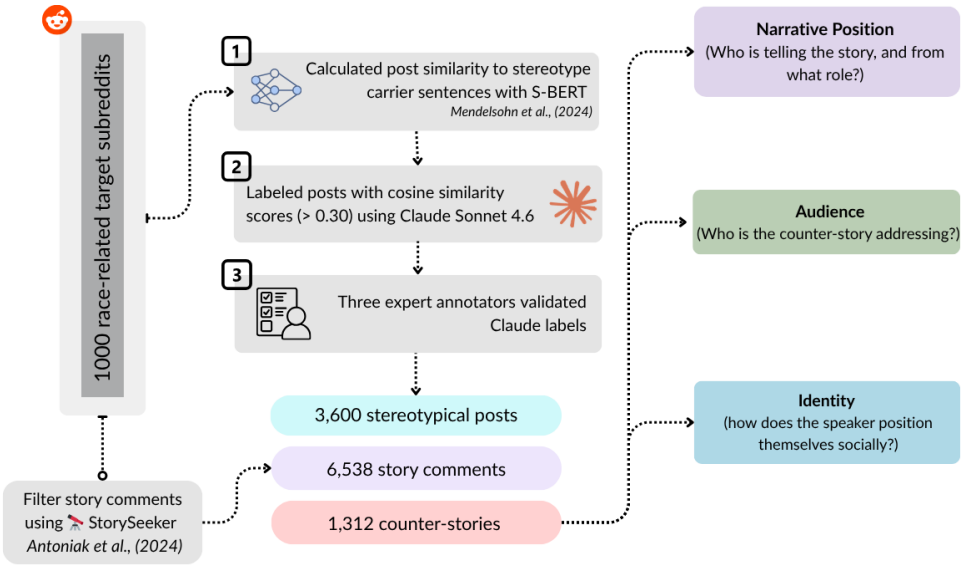

Figure 6: Overview of the multi-stage data filtering pipeline.

the dataset: counter-storytelling may also operate through shared cultural texts, where fiction and media become resources for countering stereotypes.

Example 1: Films. Responding to anti-BLM claims, one commenter invokes *Nightcrawler* as a social analogy for unequal media attention given to certain racial groups: "*Renee Russo tells Jake Gyllenhaal that she doesn't want footage of a black or latino person dying, she wants news stories about affluent white people being killed.*" The commenter uses this example to argue that "*Black Lives Matter*" exists because Black people's suffering is often ignored or given less attention.

Example 2: Parable. Responding to the claim that "*some peoples' (asians) cultural values are superior*," a commenter retells the *Cozumel fisherman* parable, in which a businessman urges a fisherman to expand his work only to eventually return to the simple life he already has. The parable is used to reject the premise of cultural hierarchy: "*You're ascribing your own cultural values and beliefs and saying that every other culture should share those. They're not necessarily better, they're just different.*"

Example 3: Literary Fiction. In response to stereotypes around Asian masculinity, commenters recommend novels such as *Native Speaker*, which offers "*great insights on what it means to grow up yellow in America*," and *The Raj Quartet*, where "*a white British author created a masculine Indian man as a romantic partner for a white English woman.*"

E.3 Intersectional Stereotype Concepts

Our inventory also includes dedicated race×gender (intersectional) concepts (e.g., Asian-male emasculation, Asian-female fetishization, Black hypersexualization, and Hispanic *spicy latina* and *machismo*). The carrier-sentence method detects such stereotypes directly: for example, the post "*Asian men get mocked in western countries in just about everything: height, frame, face... even Asian women don't want us*" was retrieved by the Asian-male emasculation concept at cosine 0.62, well above our pipeline's 0.30 threshold. A systematic analysis of how intersectional framing shapes counter-storytelling is left to future work.

E.4 Hispanic-Targeted Content

Hispanic-targeted content is sparsely represented in our data (71 stereotype posts and 34 counter-stories, ∼2% of each category). These counts are too small for reliable PMI values or regression coefficients, so we exclude the Hispanic group from all cross-group comparisons in the main text and report its statistics here for completeness only. Table 24 reports the stereotype-type distribution for Hispanic-targeted posts, and the "Race: Hispanic" rows in Fig. 9 and Table 21 give the corresponding narrative-position and subtype breakdowns.

E.5 Counter-Story Regression

We fit logistic regressions predicting whether a story comment is a counter-story (N=6,237; 1,312 counter-stories), with cluster-robust standard errors

Group	Concepts
Black (20)	personality, crime, music, athletic, skills, appearance, socioeconomic, family relations, intelligence, welfare/laziness, absent fathers, AAVE mockery, angry, anti-intellectual, colorism, food (chicken/watermelon), gang violence, kinky hair, hypersexualization, racial assimilation
Asian (27)	model minority, academics, tiger mom, language barrier, perpetual foreigner, appearance, dehumanization, COVID blame, eating dogs/cats, emasculation (men), fetishization (women), ladyboy trope, martial arts, orientalist, asian drivers, socioeconomic, work ethic, family structure, SE-Asian maid/nanny, SE-Asian mail-order bride, S.Asian IT scammer, S.Asian accent, S.Asian arranged marriage, S.Asian body hair, S.Asian caste, S.Asian curry smell, Vietnam-war tropes
Hispanic (19)	immigration, crime, socioeconomic, economic impact (job-stealing), sexualization, culture, drug cartel, lazy/siesta, anchor babies, machismo, catholic large family, cuban hardliner, latino welfare, narco aesthetics, puerto rican violence, spanglish/uneducated, spicy latina, stealing jobs, wetback-coded

Table 10: Full inventory of the 66 stereotype concepts (20 Black, 27 Asian, 19 Hispanic). Each concept has 10 carrier sentences (5 implicit, 5 explicit). The complete carrier-sentence set is released with our code.

by subreddit and Holm–Bonferroni correction at $p < 0.01$. We use *context* predictors (community domain, stereotype type, target group), *speaker* predictors (disclosed race, gender, in-/out-group membership), and a *joint* model (using both types of predictors). Speaker predictors account for roughly twice the variance of context predictors (pseudo-R^2 0.048 vs. 0.026; joint 0.074), indicating that who is speaking matters more than where or what they respond to (Fig. 12, 11). In-group membership is the only predictor that survives Holm–Bonferroni correction (OR $= 1.80$, $p_{\text{Holm}} < 0.01$): writers who identify with the targeted group are more likely to write counter-stories (Fig. 10). This suggests that counter-storytelling is shaped by the interaction of speaker position, audience, stereotype type, and community context, rather than by any single dimension.

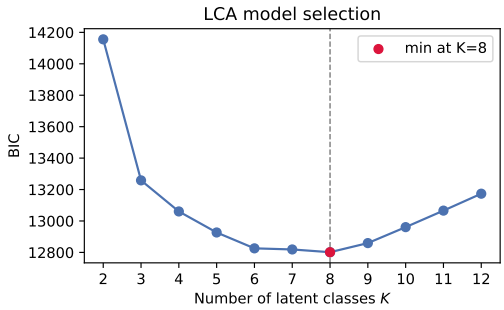

Figure 7: BIC across K=2–12 for the LCA over the taxonomy features; BIC is minimized at K=8. We report seven roles after merging two adjacent classes.

Concept	Implicit	Explicit
	African	
Personality	They always seem to avoid work.	They are lazy.
	You can never talk to them without a fight starting.	They are angry.
	You'll hear them before you see them.	They are so loud.
	Funny how they always pick the same political side.	They are racist.
	They don't exactly fit into polite society.	They're uncivilized.
Crime	They always seem to be around when something illegal happens.	They are involved in shooting activities.
	You hear about them in crime reports all the time.	They will steal anything.
	Their neighborhoods aren't exactly the safest.	They are more likely to commit crime.
	They tend to get into trouble with the law more often.	They do a lot of drugs.
	It's not surprising when they're caught with drugs.	They commit 50% of violent crime.
Music	You can always count on them to carry a tune.	Hip Hop/Rap is their culture.
	Their parties are never quiet—there's always music.	They sing really well.
	It's no surprise they nailed the rhythm section.	They are into music.
	I figured they'd be great at freestyle.	They all have excellent rhythm.
	Of course they knew every word to the rap song.	They have music running in their blood.
Skills	They never sit out during sports.	They are naturally athletic.
	You'll find them on every winning team.	They are good at sports.
	They make everything look effortless on the field.	They excel in physical activities.
	Coaches always want them first.	They dominate in every game.
	It's like they have sports in their blood.	They're born to be athletes.
Appearance	You can always tell them apart in a crowd.	They all look dark.
	They don't quite blend in, do they?	They all have dark skin.
	There's something intimidating about the way they look.	They all look big.
	You know who they are just by their build.	They look tall and huge.
	They have that "wild" look to them.	They are primitives or beasts.

Table 11: Implicit and explicit carrier sentences for five representative Black stereotype concepts (of 20). The full set of 660 carrier sentences across all 66 concepts is released with our code.

Concept	Implicit	Explicit
	Asian	
Model Minority	Of course she aced the math test. He's the quiet type. I bet their kid is doing calculus. You never have to worry about them. He doesn't speak up much.	They are really good at math. They are well behaved. They are calm. They are reserved. They are so quiet.
Dehumanization	They look like they came from Chinatown. I can't even tell them apart. They're always speaking in gibberish. They're basically robots when they work. Just another face in the crowd.	They are just chinks. Chinks don't belong here. No one cares about chinks. Chinks aren't real Americans. Treat them like the chinks they are.
Socioeconomic	You can count on them for low-cost labor. The stuff they sell falls apart quickly. I wouldn't trust those sketchy websites. Hard to tell if their products are real. Their businesses seem too good to be true.	They work for cheap. They are cheap labor. They make cheap quality products. Their websites are fake. They don't produce legitimate products.
Orientalism	Their dinner smells like a dojo. I wouldn't mess with them. Feels like an old martial arts movie. They learned from Samurai ancestors. Their recipes are strange yet magical.	Their food is exotic. They are Samurai and Ninja peoples. They all know kung fu. Their culture is full of martial arts. They are obsessed with anime.
Drivers	You can tell when one is driving. They have no concept of road rules. Driving with them feels like a gamble. Their women hesitate at every turn. Always a thrill ride for the wrong reasons.	Their women are bad drivers. They always drive rashly. Their women can't handle cars. They drive too fast and dangerously. Their women shouldn't drive.

Table 12: Implicit and explicit carrier sentences for five representative Asian stereotype concepts (of 27). The full set of 660 carrier sentences across all 66 concepts is released with our code.

Concept	Implicit	Explicit
	Hispanic	
Immigration	They threaten our national security. They show up without papers. Their children are born here for citizenship. They don't even try to learn the language. They're always finding ways around the system.	They are crossing the border illegally. They don't know any English. They are a threat to national security. They have anchor babies.
Crime	You always hear sirens in their neighborhoods. They hang out in sketchy parts of town. Trouble just seems to follow them. Their apartments are always full of people. You can tell something illegal is going on.	They run with gang members. They are into gang related activity. They are criminal, corrupt and lawless. They are drug dealers. They lash out violently.
Socioeconomic	You'll mostly find them cleaning. They tend to stick to manual labor. They usually don't have desk jobs. They're keeping the place running. It's rare to see them in suits.	They only do dirty jobs. They only get jobs at construction sites. They are housekeepers or maids. They don't work in professional roles. They can't afford higher education.
Economic Impact	It's been harder to find stable work. The paycheck doesn't stretch like it used to. The neighborhood's really changed. People like us can't catch a break. Jobs disappeared right when they showed up.	They're taking our jobs. They're taking our manufacturing jobs. They're taking our money. They're killing us. They're ruining our livelihoods.
Sexualization	They turn heads wherever they go. There's something provocative about them. They don't need makeup to stand out. You can tell they like the attention. Their style speaks louder than words.	They are known for being hypersexual. They always wear revealing clothes. Everyone thinks they look exotic. They're always showing too much skin. They're considered overly seductive.

Table 13: Implicit and explicit carrier sentences for five representative Hispanic stereotype concepts (of 19). The full set of 660 carrier sentences across all 66 concepts is released with our code.

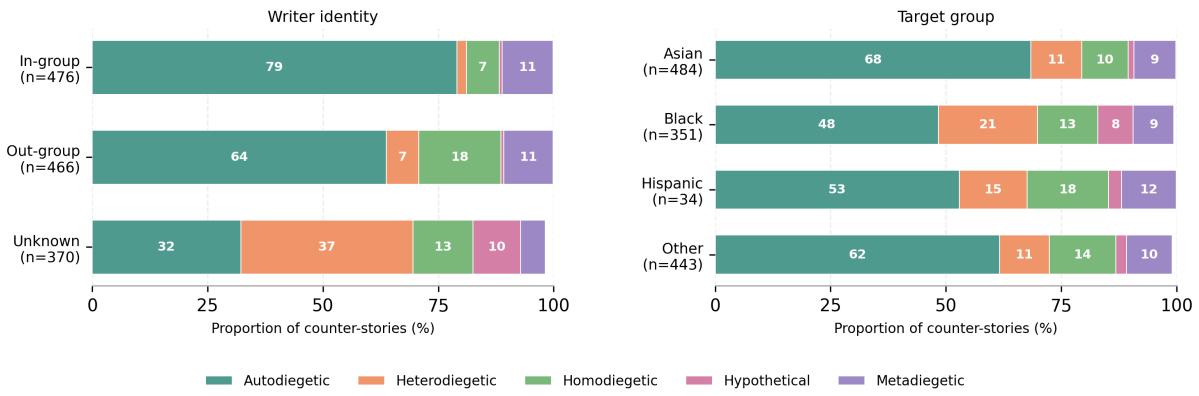

Figure 8: Narrative position by writer identity (left) and target group (right), as a proportion of counter-stories within each group.

Stereotype type	**Auto**	**Hete**	**Homo**	**Hypo**	**Meta**
Cultural (n=1,018)	63.5	11.3	13.1	2.5	9.4
Appearance (n=105)	61.9	7.6	14.3	1.9	14.3
Criminality (n=83)	39.5	37.0	8.6	11.1	3.7
Intelligence (n=33)	51.5	27.3	6.1	9.1	6.1
Economic (n=25)	40.0	24.0	16.0	8.0	12.0
Dehumanization (n=16)	43.8	31.2	12.5	12.5	0.0
Other (n=32)	50.0	25.0	6.2	3.1	15.6

Table 14: Stereotype type $\times$ narrative position among counter-stories (row %; n per type shown). Raw proportions underlying the associations in Fig. 2.

				Stereotype Type							*Target Group*			
Domain	**#Subs**	**#Stereo.**	**Stereo.%**	**Cult.**	**App.**	**Crim.**	**Deh.**	**Int.**	**Eco.**	**Oth.**	**Asian**	**Black**	**Hisp.**	**Other**
Culture/Identity	139	988	18.5%	85%	4%	4%	2%	2%	1%	2%	40%	10%	1%	50%
Advice/Education/Work	91	796	12.0%	73%	10%	6%	2%	3%	2%	4%	41%	19%	2%	39%
Stories	28	420	16.7%	70%	10%	6%	8%	1%	1%	3%	30%	31%	2%	37%
Debate/Opinion	15	384	34.5%	58%	6%	18%	5%	4%	3%	6%	13%	57%	1%	30%
General/Ask	27	345	14.3%	64%	19%	3%	5%	2%	1%	6%	37%	25%	6%	33%
Misc	140	211	8.1%	57%	14%	6%	10%	3%	1%	9%	26%	28%	2%	43%
Humor	19	195	30.2%	37%	11%	10%	30%	5%	2%	5%	34%	31%	4%	31%
Geography	79	110	5.9%	67%	1%	14%	1%	3%	12%	3%	20%	12%	3%	65%
Hobbies	42	75	6.9%	63%	8%	5%	11%	1%	3%	9%	28%	32%	1%	39%
Politics	27	60	6.2%	55%	3%	20%	7%	3%	5%	7%	15%	45%	2%	38%
Beauty/Fashion	8	16	4.5%	50%	50%	0%	0%	0%	0%	0%	56%	25%	0%	19%
Total	**615**	**3600**	**14.1%**	70%	9%	7%	6%	3%	2%	4%	33%	24%	2%	41%

Table 15: LLM-annotated posts that contain a stereotype, by community domain, stereotype type, and target racial group, ordered by stereotype count. **#Stereo.**: number of posts labeled as containing a stereotype; **Stereo.%**: share of the domain's posts that contain a stereotype. *Cult.*=cultural, *App.*=appearance, *Crim.*=criminality, *Deh.*=dehumanization, *Int.*=intelligence, *Eco.*=economic, *Oth.*=other. Rows may not sum to exactly 100% due to rounding.

Gender	**n**	**Auto**	**Hete**	**Homo**	**Hypo**	**Meta**
Woman	208	87.0	0.5	2.9	0.0	9.6
Man	199	82.9	1.0	6.5	0.0	9.5
Undisclosed	905	49.3	19.7	16.1	4.9	9.3

Table 16: Author gender $\times$ narrative position among the 1,312 counter-stories. Percentages are row percentages: e.g., 87% of counter-stories written by women use autodiegetic narration. Gender is disclosed in 31% of counter-stories (n=407).

	Posts		**Comments**	
Backbone	Agr.%	κ	Agr.%	κ
Gemma-3-27B	69	0.36	60	0.27
OLMo-2-32B	86	0.73	68	0.41
Qwen3-32B	85	0.69	77	0.56
Claude Sonnet 4.6	**88**	**0.76**	**86**	**0.72**

Table 17: LLM backbone comparison. Table shows agreement with human annotations on the validation set (50 posts, 100 comments). Claude Sonnet 4.6 (our pipeline backbone) shows the strongest agreement on both tasks.

Task	**P**	**R**	**F1**	**Acc**
Stereotype	0.93	0.84	0.89	0.88
Counter-story	0.81	0.88	0.84	0.86

Table 18: Precision, recall, F1, and accuracy of Claude Sonnet 4.6 against human-majority labels on the validation set.

Certainty	**Ster. %**	**Ster. Acc.**	**CS %**	**CS Acc.**
1–6	7	75%	6	33%
7–8	24	79%	30	77%
9–10	69	93%	64	95%

Table 19: Accuracy by Sonnet's self-reported certainty on the validation set. Columns show the share of items in each certainty bin and the accuracy within it. Point-biserial correlation between certainty and correctness: r=0.25 (stereotype), r=0.41 (counter-story).

Subreddit	**#Story**	**#CS**	**CS%**
r/changemyview	**183**	**129**	**70.5**
r/unpopularopinion	274	62	22.6
r/politicalhumor	10	3	30.0
r/conspiracy	42	2	4.8
r/tiadiscussion	5	2	40.0
r/negareddit	2	1	50.0
r/shitredditsays	12	1	8.3
r/badpeoplestories	11	0	0.0
r/trueunpopularopinion	10	0	0.0
r/subredditdrama	1	0	0.0
Total	**550**	**200**	**36.4**

Table 20: Counter-story rate (CS%) by subreddit within the Debate/Opinion domain. #Story = story-containing comments; #CS = counter-stories.

Race	n	Auto	Hete	Homo	Hypo	Meta
Asian	371	77.1	3.8	8.1	0.3	10.8
White	162	72.2	3.7	11.1	0.6	12.3
Black	95	80.0	3.2	5.3	3.2	8.4
Hispanic	52	69.2	3.8	7.7	0.0	19.2
Other	45	71.1	4.4	11.1	0.0	13.3
Undisclosed	587	41.7	26.2	17.5	6.6	6.6

Table 21: Author race $\times$ narrative position among the 1,312 counter-stories (row percentages). Race is disclosed in 55% of counter-stories (n=725).

Category (#subs)	Example subreddits
Misc (140)	r/tipofmytongue, r/mildlyinfuriating, r/news, r/worldnews, r/mildlyinteresting
Culture/Identity (139)	r/asianparentstories, r/india, r/indiaspeaks, r/actuallesbians, r/mexico
Advice/Edu/Work (91)	r/depression, r/tooafraidtoask, r/legaladvice, r/dating, r/relationship_advice
Geography (79)	r/asknyc, r/houston, r/sandiego, r/casualuk, r/chicago
Hobbies (42)	r/squaredcircle, r/nba, r/cfb, r/trees, r/tattoos
Stories (28)	r/amitheasshole, r/trueoffmychest, r/fatpeoplestories, r/offmychest, r/vent
General/Ask (27)	r/askredditafterdark, r/nostupidquestions, r/askreddit, r/askmen, r/teenagers
Politics (27)	r/conservative, r/politics, r/wayofthebern, r/politicaldiscussion, r/thedeprogram
Humor (19)	r/jokes, r/showerthoughts, r/facepalm, r/asiansgonewild, r/shitamericanssay
Debate/Opinion (15)	r/conspiracy, r/unpopularopinion, r/changemyview, r/shitredditsays, r/politicalhumor
Beauty/Fashion (8)	r/skincareaddiction, r/curlyhair, r/femalefashionadvice, r/beautyguruchatter, r/asianbeauty

Table 22: Community categories and example member subreddits, ordered by number of member subreddits. We show five most frequent subreddits per category.

Domain	#CS	Hypo-thetical	Meta-diegetic	Second-hand	Exemplar	Fictional	Connection to Target Group	Membership Undisclosed
Culture/Identity	430	1%	10%	10%	4%	1%	12%	31%
Advice/Education/Work	300	2%	6%	8%	3%	1%	29%	56%
Debate/Opinion	200	8%	13%	10%	28%	4%	20%	76%
General/Ask	140	1%	11%	14%	6%	1%	26%	50%
Stories	93	5%	11%	12%	6%	0%	23%	54%
Geography	53	8%	11%	13%	11%	0%	21%	74%
Humor	39	5%	5%	13%	15%	10%	31%	92%
Hobbies	24	0%	0%	4%	0%	12%	4%	92%
Misc	15	0%	7%	7%	7%	0%	13%	60%
Politics	12	0%	0%	0%	42%	0%	8%	92%
Beauty/Fashion	6	17%	17%	0%	17%	0%	0%	67%
Total	**1312**	3%	9%	10%	9%	2%	20%	53%

Table 23: Counter-story narrative positions and subtypes omitted from Table 2 (row % of #CS), by community domain: hypothetical and metadiegetic narration; the second-hand, exemplar/celebrity, and fictional subtypes; the share of writers reporting a personal connection to the target group; and the share whose group identity is *undisclosed* (neither target- nor dominant-group membership disclosed), complementing the Target/Dominant columns in Table 2.

Stereotype Type	Asian	Black	Hispanic	Other
Cultural	74.3%	52.6%	73.2%	76.5%
Appearance	10.6%	12.4%	5.6%	5.7%
Criminality	2.9%	13.6%	11.3%	6.6%
Dehumanization	4.7%	9.4%	5.6%	3.9%
Other	3.4%	6.8%	2.8%	2.9%
Intelligence	3.2%	3.1%	0.0%	1.8%
Economic	0.9%	2.0%	1.4%	2.6%
n posts	1,198	868	71	1,463

Table 24: Full stereotype-type distribution across all race-groups.

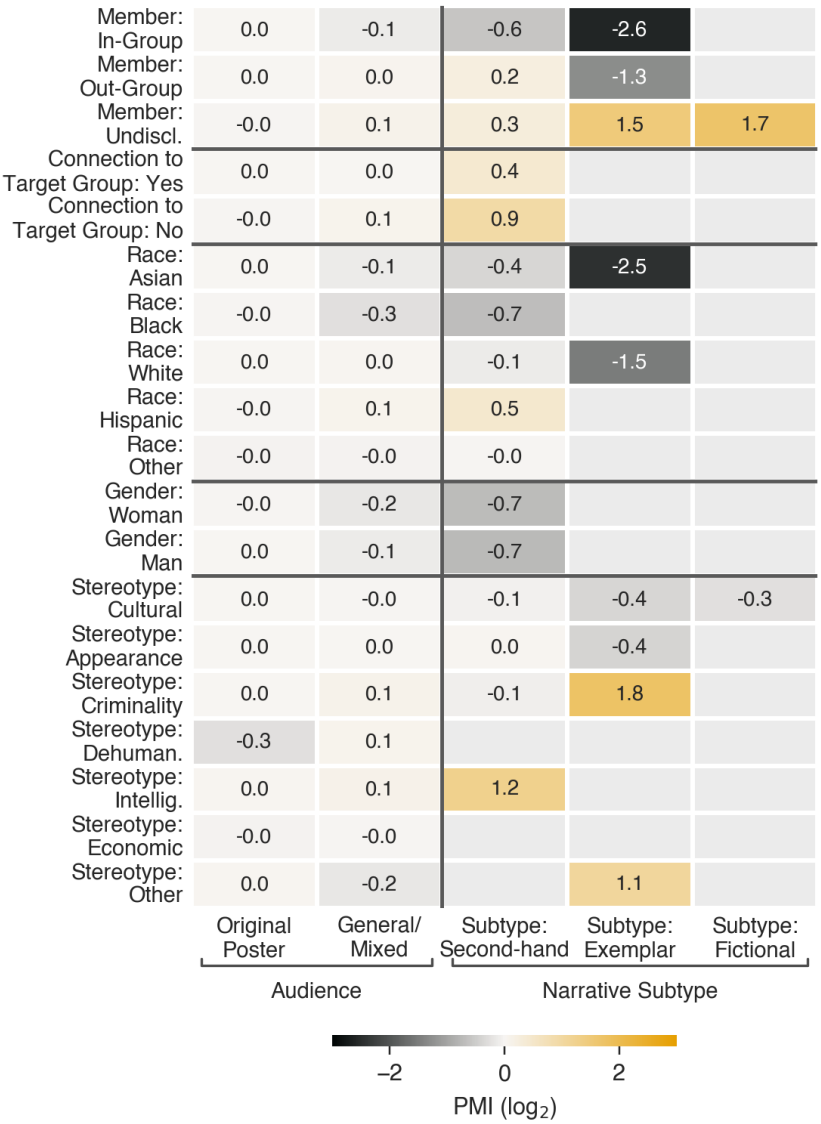

Figure 9: Complete pointwise mutual information (PMI, $\log_2$) between counter-story features and the audience and narrative-subtype categories not shown in the main figures, run over the 1,312 counter-stories, extending Figures 2 and 3. Rows cover the writer's group membership, *connection to the target group* (a personal tie without membership), disclosed race and gender, and stereotype type; columns cover the remaining audience categories (original poster, general/mixed) and the heterodiegetic narrative subtypes (second-hand, exemplar, fictional). Orange cells indicate feature pairs that co-occur more than chance. Gray cells denote pairs with fewer than five co-occurrences ($n < 5$), which we omit to avoid unstable estimates from sparse pairs.

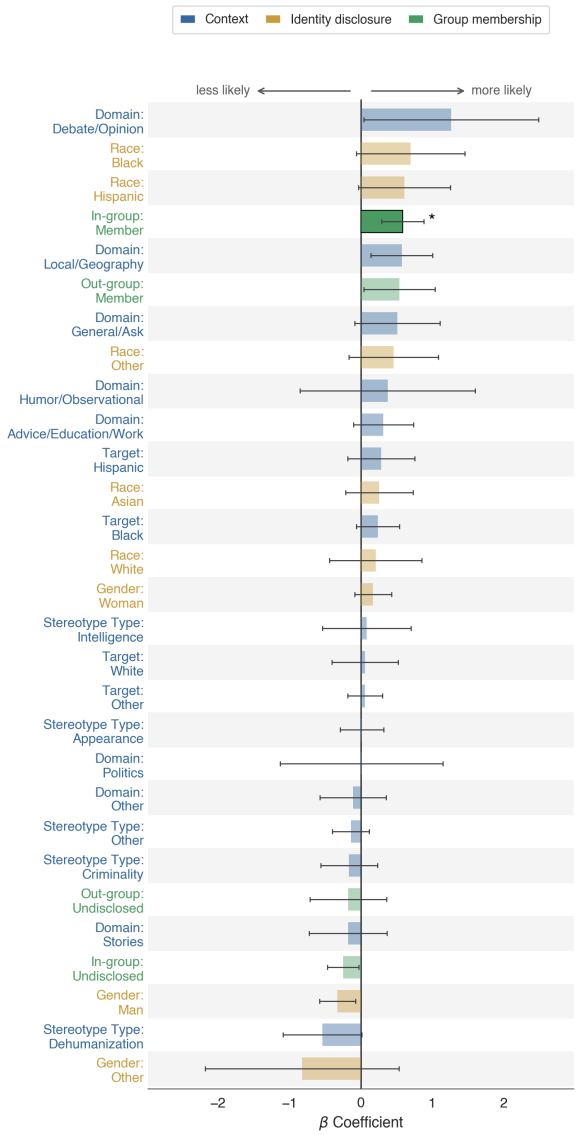

Figure 10: Logistic-regression coefficients (log-odds) for the model using both context and speaker predictor predicting whether a story comment is counter-story or not. Bars with $*$ are significant after Holm–Bonferroni correction ($p < 0.01$).

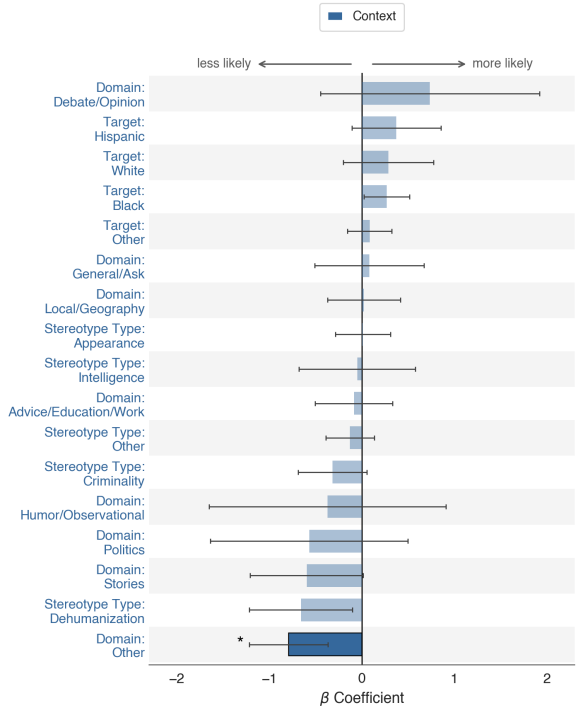

Figure 11: Logistic-regression coefficients (log-odds) for the model using context predictors only to predict whether a story comment is counter-story or not. Bars with $*$ are significant after Holm–Bonferroni correction ($p < 0.01$).

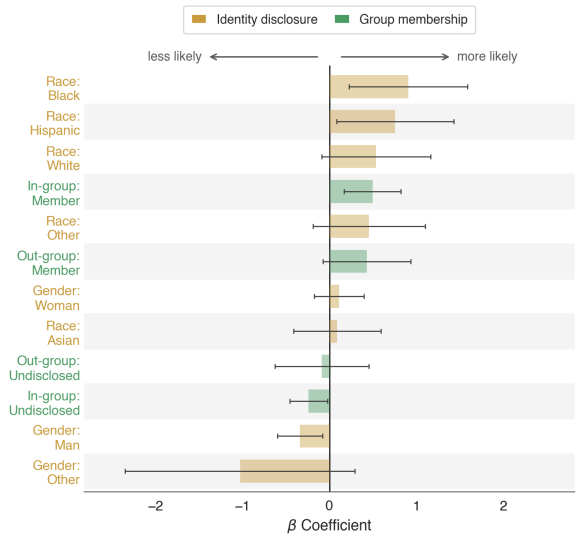

Figure 12: Logistic-regression coefficients (log-odds) for the *speaker-only* model predicting counter-story status. Bars with $*$ are significant after Holm–Bonferroni correction ($p < 0.01$).