Detecting Rhetorical Distortion Across Platforms: A Multi-Platform, Multi-Language Classifier for Influencer Content Analysis

September 3, 2026

Abstract

As social media platforms compete for attention, content creators increasingly rely on rhetorical distortion—defined as the systematic departure from rhetorical proportionality in linguistic framing [5, 1]—to maximize engagement. Existing computational approaches to online manipulation target either factual inaccuracy [16], clickbait headlines [13, 3], or coordinated propaganda [4]; none of these frameworks is designed to capture content that is factually accurate yet rhetorically disproportionate to its underlying evidential basis—the dominant manipulation strategy in influencer content. We close this gap with a literature-grounded, three-tier classification system for detecting five distortion dimensions—Significance Inflation (SI), Anxiety Manufacturing (AM), Novelty Claims (NC), Loaded Language (LL), and Temporal Distortion (TD)—each independently derived from established theories in rhetoric, persuasion research, and propaganda detection, so that every scored dimension is traceable to a specific theoretical mechanism rather than an ad hoc heuristic. The system combines rule-based keyword matching, negative-pattern filtering, and GPT-4o-mini bidirectional verification, applied to 332 accounts and 13,219 posts across six platforms spanning English and Chinese content.

Each dimension is reported independently as the primary measure. A composite Distortion Index (DI), computed as the unweighted arithmetic mean of the five dimension rates, serves as an auxiliary summary indicator. A Consistency Score (CS) is reported separately to indicate the temporal stability of each account's distortion pattern; accounts with CS < 0.30 are flagged as low-confidence. We validate the system against a 200-post human-annotated benchmark (Cohen's $\kappa = 0.76$), where the full three-tier pipeline improves precision by 12 percentage points over rule-only classification ($0.71 \rightarrow 0.83$), demonstrating that the GPT-4o-mini verification stage is a necessary—not merely convenient—component of the design.

Our central finding is that distortion patterns are *platform-mediated* rather than content-determined: Twitter/X (mean DI $= 4.7$) and YouTube (mean DI $= 2.9$) substantially exceed long-form platforms (RSS mean DI $= 0.8$, Bluesky mean DI $= 0.9$). Loaded Language is the dominant dimension on Twitter/X (LL $= 8.4\%$), while Significance Inflation leads on YouTube (SI $= 4.8\%$). Our results suggest that platform affordances, not individual creator intention, are the primary driver of rhetorical distortion in online influencer content.

1 Introduction

Problem. The rise of social media influencers as primary information intermediaries has introduced new forms of rhetorical manipulation into everyday information consumption. Unlike traditional journalistic gatekeeping, influencer content is governed by engagement metrics rather than editorial standards, creating structural incentives for exaggeration, emotional amplification, and false urgency. Consider a financial commentator who accurately reports that "markets fell 0.3%" but frames it as "CATASTROPHIC COLLAPSE you need to act NOW." No individual fact in this post is false, so the post is invisible to fact-checking pipelines;

yet the framing systematically overstates the evidential weight of the underlying event, which is precisely the kind of manipulation that shapes what audiences believe matters. Understanding how this class of distortion manifests across platforms and languages is essential for information literacy research, platform governance, and computational fact-checking.

Gap in prior work. Prior work on misinformation detection has focused primarily on factual inaccuracies [16], clickbait headlines [13, 3], and coordinated propaganda [4]. These approaches treat distortion as a property of content truth-value (is the claim true or false) or of a single structural cue (does the headline withhold information), leaving unexamined a broader category of rhetorical manipulation that is technically truthful, standalone, and yet misleadingly disproportionate to the evidence it is built on. A second, independent gap is theoretical: existing computational systems for detecting persuasive or manipulative language [14, 10, 8] typically define their target categories operationally, from observed corpus patterns, rather than by grounding each category in an independently motivated theory of rhetoric or persuasion—which makes the resulting categories difficult to defend against the objection that they are merely descriptive artifacts of one dataset.

Our approach. We address both gaps with a multi-platform, multi-language system that detects five rhetorical distortion dimensions, each independently grounded in established scholarship *before* being operationalized into detection rules: Significance Inflation (SI; Entman 5), Anxiety Manufacturing (AM; Witte 19, Witte and Allen 20), Novelty Claims (NC; Da San Martino et al. 4, Flanagin and Metzger 6), Loaded Language (LL; Da San Martino et al. 4, Weston 18), and Temporal Distortion (TD; Wang and Strong 17). These five dimensions operationalize the Aristotelian triad of rhetorical appeals [1] as reframed through computational persuasion research [12]: SI and TD distort *logos* (logical appeal); AM and LL distort *pathos* (emotional appeal); NC distorts *ethos* (credibility appeal). Each dimension is detected by a three-tier pipeline (rule-based matching, negative-pattern filtering, and bidirectional LLM verification, §3.5) and reported independently, so that no single opaque composite score stands in for the underlying evidence. Our system was applied to 332 accounts across six platforms, generating 13,219 classified posts, and validated against a human-annotated benchmark to demonstrate that the added verification stage is empirically, not just conceptually, necessary (§4.6).

Contributions.

1. A theoretically grounded five-dimension framework for rhetorical distortion, with each dimension derived from peer-reviewed scholarship in rhetoric, persuasion research, and propaganda detection.
2. A three-tier classification pipeline combining rules, negative filtering, and GPT-4o-mini bidirectional verification, with an equal-weight composite DI and a separate Consistency Score.
3. A dataset of 332 accounts and 13,219 posts across six platforms and two languages.
4. Evidence that platform affordances—not creator demographics or content category—are the strongest predictor of distortion level, with Twitter/X showing 5.9$\times$ higher mean DI than RSS/Newsletter.

Research Questions. This study is governed by the following research questions:

RQ1 *Platform effect:* To what extent do platform affordances predict the level and type of rhetorical distortion in influencer content, independent of content category and creator identity?

RQ2 *Dimension structure:* How do the five distortion dimensions (SI, AM, NC, LL, TD) distribute across platforms, and do platforms exhibit distinct characteristic distortion signatures?

RQ3 *Cross-linguistic validity:* To what extent are rhetorical distortion patterns generalizable across English and Chinese-language social media platforms?

RQ4 *Amplification:* Does community-generated content (replies and comments) amplify rhetorical distortion beyond the level present in creator-generated posts?

These questions address gaps identified in prior work on platform affordances [2], cross-linguistic NLP generalizability [11, 15], and community-level distortion amplification, none of which has been examined systematically in the context of rhetorical distortion in multi-platform influencer content.

2 Background and Related Work

2.1 Computational Approaches to Rhetorical Manipulation

Early work on automated detection of persuasive language focused on political propaganda [14] and emotionally charged news [10]. These studies established that lexical features—emotional valence, certainty markers, generalization language—provide reliable signals for manipulation detection. However, they were limited to political news corpora and did not address the broader space of social media influencer content.

Clickbait detection has received sustained attention [13, 3], with approaches ranging from lexical heuristics to neural classifiers. The emergence of large language models has opened new possibilities for nuanced rhetorical analysis: Goldstein et al. [8] demonstrated that GPT-4 can detect subtle persuasion techniques with near-human accuracy when provided with appropriate context. Our system uses GPT-4o-mini in a verification role rather than as the primary classifier, reducing API costs while maintaining accuracy through bidirectional false-positive and false-negative checking.

2.2 Platform Effects on Content Production

The concept of *platform affordances* refers to the action possibilities that a platform's technical design, algorithmic structure, and social norms make available to users—and, by extension, the constraints it places on content production [2]. Affordances relevant to rhetorical distortion include character limits (incentivizing compressed, emotionally salient language), algorithmic amplification (rewarding engagement-maximizing rhetoric), audience feedback loops (normalizing high-distortion styles that receive visible positive responses), and monetization structures (creating financial incentives for attention-capturing framing). Critically, platform affordances shape content characteristics *independent of creator identity*: the same person produces measurably different rhetorical styles across platforms because the affordance structures differ [7]. Cross-platform studies consistently find that content characteristics diverge by platform even when creator identity is held constant, establishing platform affordances as a structural rather than individual-level predictor of rhetorical style.

2.3 Cross-Linguistic Distortion Patterns

Comparative analysis of Chinese and English social media has documented systematic differences in information style, emotional expression, and authority appeals [11, 15]. Chinese online media culture has developed distinct genres of financial and technology commentary characterized by highly compressed, impact-maximizing language. Our inclusion of Weibo allows us to test whether our classifier, developed primarily from English-language patterns, generalizes to Chinese content or requires language-specific calibration.

Key differences from prior work. Unlike misinformation detection [16], which requires an external ground-truth judgment of factual accuracy, our system requires no fact-verification step: distortion is defined and measured entirely at the level of linguistic framing, so a post can be flagged even when every underlying fact it reports is correct. Unlike clickbait detection [13, 3], which targets a single structural cue (a headline that withholds information to induce a click), our five dimensions jointly cover the three classical rhetorical appeals (*logos*, *pathos*, *ethos*) and are each independently grounded in a distinct body of scholarship rather than a single detectable pattern. Unlike general-purpose propaganda taxonomies [4], which were developed for professionally edited news articles, our pipeline is adapted to the short-form, multilingual, high-volume setting of influencer social media, and adds an explicit temporal-stability metric (the Consistency Score, §3.2) that propaganda-detection taxonomies do not provide. In summary, we shift the unit of analysis from *is this claim true* or *is this headline structurally misleading* to *is this framing proportionate to the evidence*, which is a category of harm that existing detection frameworks are not designed to capture.

3 Methodology

3.1 Problem Definition

We formalize account-level distortion profiling as follows. Let a denote an account with a post history $P(a) = \{p_1, p_2, \ldots, p_n\}$, where each post p_i is associated with a timestamp t_i. For each of the five distortion dimensions $d \in \{\mathrm{SI}, \mathrm{AM}, \mathrm{NC}, \mathrm{LL}, \mathrm{TD}\}$ (defined and grounded individually in §3.2), the three-tier pipeline C (§3.5) maps each post to a binary indicator,

$$C_d(p_i) \in \{0, 1\}, \tag{1}$$

where $C_d(p_i) = 1$ if post p_i triggers dimension d. The account-level *dimension rate* is the empirical trigger frequency over the post history:

$$\mathrm{Rate}_d(a) = \frac{1}{n} \sum_{i=1}^{n} C_d(p_i). \tag{2}$$

The five dimension rates $\{\mathrm{Rate}_d(a)\}_d$, reported independently, are the primary outcome of the system (*cf.* Eq. 3 for the auxiliary composite index and Eq. 4 for the companion stability metric). This formulation deliberately departs from an external-ground-truth formulation such as $\Pr[\text{claim in } p_i \text{ is false}]$: because C_d operates only on the linguistic form of p_i, no fact-verification oracle is required, which is the key property that distinguishes this task from misinformation detection (§2).

3.2 Theoretical Framework

Core Definition. We define *Rhetorical Distortion* as the systematic, measurable disproportion between a communicator's linguistic framing of information and the evidential basis of that information [9]. This operationalizes the broader concept of *rhetorical proportionality* [5]: the degree to which framing reflects the actual evidential strength, importance, urgency, novelty, and emotional implications of the underlying content.

We ground this definition in Aristotle's [1] three rhetorical appeals—*logos* (logical argument), *pathos* (emotional appeal), and *ethos* (credibility)—as operationalized for computational analysis by Pauli et al. [12]. Rhetorical distortion occurs when these appeals are systematically *disproportionate* to the content they frame.

Five Distortion Dimensions. We operationalize rhetorical distortion through five dimensions, each grounded in a distinct body of scholarship:

1. **Significance Inflation (SI)** — *Logos distortion.* SI refers to the systematic exaggeration of the importance or magnitude of information beyond its evidential support [5]. Entman's framing theory identifies *salience manipulation*—making certain aspects of information more salient than the evidence warrants—as the core mechanism of this distortion type. Kreider [9] further grounds SI in informal logic, classifying it as an argumentative fallacy of proportionality. Representative patterns: "this changes everything," "once in a generation."

2. **Anxiety Manufacturing (AM)** — *Pathos distortion.* AM refers to the construction of fear or urgency targeting the reader's career or personal adequacy, in excess of what evidence warrants. This operationalizes the fear appeal component of Witte's [19] Extended Parallel Process Model (EPPM), which identifies threat severity and personal susceptibility as the two core elements of fear-based persuasion. Witte and Allen's [20] meta-analysis confirms that fear appeals targeting personal vulnerability produce the largest effect sizes of any persuasion mechanism. Representative patterns: "you will be left behind," "unemployable in 6 months."

3. **Novelty Claims (NC)** — *Ethos distortion.* NC refers to false assertions of epistemic priority or exclusivity: claiming to be first, to have discovered something hidden, or to have access unavailable to others [6]. Da San Martino et al. [4] classify this as the "Appeal to Authority" propaganda technique when the claimed authority is false or unwarranted. Representative patterns: "nobody is talking about this," "I discovered a secret technique."

4. **Loaded Language (LL)** — *Pathos distortion.* LL refers to the use of emotionally charged words or phrases whose affective intensity exceeds what the underlying content warrants [4]. Weston [18] systematizes loaded language as a rhetorical fallacy in which emotional vocabulary substitutes for substantive argument. LL is the most frequent propaganda technique in the Da San Martino et al. corpus, appearing in sensational adjectives and adverbs, excessive capitalization, and repeated exclamation marks. Representative patterns: "catastrophic," "devastating collapse," ALL-CAPS emphasis, "!!!"

5. **Temporal Distortion (TD)** — *Logos distortion.* TD refers to the misrepresentation of information recency: framing old content as breaking, urgent, or newly discovered. This operationalizes the *timeliness* dimension of Wang and Strong's [17] information quality framework, which identifies temporal accuracy as a fundamental dimension of information quality distinct from factual accuracy. Representative patterns: "breaking news" applied to content weeks old, "just dropped" for previously published material.

Distortion Index. At the account level, we compute five dimension rates (fraction of posts triggering each dimension) as the primary outcome measures. A composite Distortion Index (DI) is computed as the unweighted arithmetic mean of the five rates:

$$\text{DI} = \frac{SI + AM + NC + LL + TD}{5} \times 100 \tag{3}$$

Equal weighting follows the convention of multi-dimensional information quality frameworks that assign equal weight in the absence of empirical evidence for differential weighting [17]. Each dimension is reported independently as the primary measure; DI serves only as an auxiliary summary for cross-platform comparison.

Consistency Score. A Consistency Score (CS) measures the temporal stability of each account's distortion pattern. Posts are binned by calendar week; CS is derived from the standard deviation of weekly distortion rates:

$$CS(a) = \max(0,\ 1 - 2\sigma_w), \tag{4}$$

where σ_w is the standard deviation of weekly distortion rates. CS $= 1.0$ indicates a perfectly stable distortion pattern; CS $= 0.0$ indicates high week-to-week variability. Accounts with CS < 0.30 are flagged as low-confidence and excluded from cross-platform comparisons, as their distortion patterns may reflect episodic rather than systematic behavior. CS is reported separately and does not contribute to DI.

3.3 Platform and Account Selection

We selected six platforms representing the major social media modalities: RSS/Newsletter (long-form asynchronous), YouTube (video with transcripts), Bluesky (English short-form social), Reddit (community discussion), Weibo (Chinese short-form social), and Twitter/X (English micro-blog with comments). Within each platform, accounts were selected to cover multiple content categories while including a range of expected distortion levels for validity testing.

Table 1: Corpus overview across six platforms. Twitter/X classification text = tweet content merged with top-5 reply comments; DI reflects both creator posts and community reply content.

Platform	**Accts**	**Posts**	**Mean DI**	**Max DI**	**Lang.**	**Format**
RSS/Newsletter	66	1,262	0.8	10	EN	Title + abstract (800c)
YouTube	32	640	2.9	10	EN	Title + transcript (800c)
Bluesky	54	2,680	0.9	5	EN	Full post (300c)
Reddit	58	2,790	0.8	4	EN	Title + body (300c)
Weibo	69	3,240	0.3	3	ZH	Full post (2000c)
Twitter/X	53	2,607	4.7	13	EN	Tweet + comments (1000c)
Total	**332**	**13,219**	—	**13**	EN/ZH	—

3.4 Data Collection

For each platform we used different collection strategies adapted to platform APIs and access restrictions. RSS/Newsletter and YouTube used public RSS feeds. Bluesky was collected via the public ATP protocol API. Reddit was collected via Playwright browser automation following the platform's 2023 API restriction changes; each post's selftext was fetched from the individual post JSON endpoint to supplement titles. Weibo was collected via the mobile API (`m.weibo.cn`) after Playwright initialization of a Visitor Session. Twitter/X used Playwright-based collection with authenticated session cookies, incorporating virtual scroll collection to address DOM virtualization.

3.5 Three-Tier Classification Pipeline

Stage 1: Rule-Based Keyword Detection. We define two signal tiers for each of five distortion dimensions (see §3.2). Strong signals (one suffices to trigger) have weight STRONG_WEIGHT $= 0.35$; weak signals (two or more required) have weight WEAK_WEIGHT $= 0.15$. Confidence accumulates additively, capped at CONFIDENCE_CAP $= 0.95$.

Table 2: Five distortion dimensions with representative signal patterns. Dimension grounding references are in §3.2.

Dimension	Strong signals (1 triggers)	Weak signals (≥2 trigger)
SI: Significance inflation	"changes everything", "once in a generation", "never seen anything like this"	"most important", "revolutionary", "unprecedented", "game changer"
AM: Anxiety manufacturing	"you will be left behind", "unemployable in 6 months", "running out of time"	"too late", "falling behind", `(threat|crisis){0,20}(career|job)`
NC: Novelty claims	"nobody is talking about this", "I've been warning about this for years"	"I (found|discovered|built)", "secret/hidden truth"
LL: Loaded language	"devastating", "catastrophic", "mind-blowing", `(killing|destroying){0,20}(industry|career)`	"shocking", "brutal", "jaw-dropping", `[A-Z]{4,}, !{2,}`
TD: Temporal distortion	"breaking news", "just dropped", "this just in"	"just announced/released", "happening now"

Chinese-language content uses a parallel keyword set including: 震惊/骇人/毁灭性 (loaded language), 重磅/史上/里程碑 (significance inflation), 危机/崩溃/恐慌 (anxiety manufacturing), 首次/独家/曝光 (novelty claims), and 突发/刚刚 (temporal distortion). Negative filters for Chinese content additionally exclude 历史上的今天 (temporal: historical retrospectives), 官方发布 (novelty: official announcements), and 不要慌/理性看待 (anxiety: calming language).

Stage 2: Negative Pattern Filtering. Negative filtering applies regex patterns to exclude known false-positive contexts. Key exclusions include: (a) temporal dimension: "this (morning|week)" followed by first-person activity verbs, identifying personal scheduling language rather than breaking news; (b) anxiety dimension: "never too late to (learn|start|improve)" identifying encouragement rather than fear-mongering; (c) novelty dimension: content inside quotation blocks where the author is attributing claims to others. Negative filtering is intentionally conservative—it removes only high-confidence false positives, leaving borderline cases for GPT-4o-mini verification. This two-stage design reduces API costs by approximately 30% compared to submitting all flagged posts to the LLM.

Stage 3: GPT-4o-mini Bidirectional Verification. Three paths exist after Stages 1–2.

Path A: Low-confidence path (conf < 0.70)

`classify_llm_fresh()`: GPT-4o-mini classifies from scratch without any prior rule output, preventing confirmation bias from Stage 1.

Path B: Verification path (conf ≥ 0.70, VERIFY_ALL=True)

`verify_with_gpt()`: bidirectional checking in one API call. (1) False-positive scan: are Stage 1's flagged dimensions genuine? (2) False-negative scan: were dimensions excluded by Stage 2 negative filtering incorrectly excluded?

Path C: Direct return (no Stage 1 or Stage 2 hits)

Post is returned with `confidence` $= 1.0$ and empty distortion types. No API call required.

The GPT-4o-mini system prompt instructs the model to return structured JSON with fields `types`, `confidence`, `signals`, and `corrections`. All GPT calls use `temperature` $= 0$ for reproducibility.

3.6 Distortion Index and Consistency Score

The Distortion Index (DI) and Consistency Score (CS) are defined in §3.2 (Equations 3 and 4). Each of the five dimension rates is reported independently as the primary outcome; DI serves as an auxiliary summary for cross-platform comparison only. Accounts flagged as low-confidence (CS < 0.30) are excluded from cross-platform analyses: in our corpus, 11 accounts across all six platforms met this criterion (RSS: *timferriss*; YouTube: *minoritymindset*; Reddit: *u/GallowBoob*; Weibo: one account; Twitter: 7 accounts including *realDonaldTrump*, *WHO*).

4 Results

4.1 Finding 1: Platform Format Predicts Distortion Level (RQ1)

The most striking finding is the consistent separation between Twitter/X (mean DI $= 4.7$, SD $= 3.4$) and all other platforms (mean DI range 0.3–2.9), with RSS/Newsletter and Reddit showing the lowest average distortion (both mean DI $= 0.8$).

Table 3: Platform-level distortion indices and dimension breakdown (averaged across all accounts per platform; DI = unweighted mean of five rates).

Platform	n	Mean DI	SD	Max	SI%	AM%	NC%	LL%
RSS/Newsletter	66	0.8	1.7	10	1.5	0.4	0.4	0.0
YouTube	32	2.9	3.1	10	4.8	2.5	2.0	2.5
Bluesky	54	0.9	1.0	5	1.1	0.9	0.4	1.7
Reddit	58	0.8	1.0	4	1.3	1.3	0.2	0.8
Weibo	69	0.3	0.6	3	0.5	0.4	0.3	0.3
Twitter/X	53	4.7	3.4	13	6.4	5.3	1.6	8.4

Twitter/X is the highest-distortion platform across all five dimensions, with Loaded Language (LL $= 8.4\%$) as its dominant dimension. This reflects the merged tweet-plus-comment classification unit: Twitter/X posts were classified together with their top-5 replies, capturing the distortion amplification that occurs in reply threads.

YouTube (mean DI $= 2.9$) is the highest-distortion platform excluding Twitter/X. Significance Inflation is the leading dimension (SI $= 4.8\%$), consistent with the production incentive to frame video content as significant to justify viewer time investment.

Weibo (mean DI $= 0.3$) shows the lowest composite DI among all platforms, despite being a short-form Chinese social platform. This reflects both the composition of our Weibo sample (dominated by official media and institutional accounts with low inherent distortion) and the limitation that our rule-based system has lower coverage of Chinese-language distortion patterns than English-language ones.

RSS/Newsletter and Reddit (both mean DI $= 0.8$) cluster together as the most consistent low-distortion platforms, supporting the view that long-form publication cycles and community moderation norms both function as distortion suppressants.

4.2 Finding 2: Loaded Language Dominates Twitter/X; Anxiety Manufacturing Differentiates Social Platforms (RQ2)

Among the five distortion dimensions, Loaded Language (LL) shows the most dramatic platform concentration. Twitter/X LL $= 8.4\%$ dwarfs all other platforms (RSS $= 0.0\%$, Weibo $= 0.3\%$, Reddit $= 0.8\%$, Bluesky $= 1.7\%$, YouTube $= 2.5\%$), a ratio of effectively ∞ against RSS/Newsletter. This concentration reflects both the rhetorical affordances of the Twitter/X format (short posts incentivize emotionally salient language) and the comment-merging methodology, which captures emotionally charged user responses.

Anxiety Manufacturing (AM) shows a complementary pattern among the non-Twitter platforms. RSS/Newsletter (AM $= 0.4\%$) and Weibo (AM $= 0.4\%$) have the lowest AM rates, while Twitter/X (AM $= 5.3\%$) leads substantially. The highest AM account in our corpus is `@Reuters` on Twitter/X (AM $= 18\%$), reflecting that breaking-news wire framing systematically activates anxiety-adjacent language ("urgent," "developing," "immediate") even in journalistically neutral content. This case illustrates how platform-format constraints produce distortion-adjacent language independent of editorial intent.

4.3 Finding 3: Significance Inflation Leads on Twitter/X and YouTube; Cross-Linguistic Patterns Require Caution (RQ2, RQ3)

Significance Inflation (SI) shows the clearest platform gradient after Twitter/X: YouTube (SI $= 4.8\%$) and Twitter/X (SI $= 6.4\%$) substantially exceed RSS/Newsletter (SI $= 1.5\%$), Bluesky (SI $= 1.1\%$), Reddit (SI $= 1.3\%$), and Weibo (SI $= 0.5\%$).

Table 4: Significance Inflation rates for Chinese vs. English accounts within the same content category (AI/ML and finance). Caution: lower Weibo SI may reflect rule coverage limitations rather than genuine lower distortion.

Category	Language/Platform	Accounts	SI Rate
AI/ML	Chinese (Weibo)	5	0.8%
AI/ML	English (all EN plat.)	29	2.0%
Finance	Chinese (Weibo)	19	0.9%
Finance	English (all EN plat.)	34	4.3%

Counter-intuitively, Weibo SI rates are lower than English-platform rates within the same categories. This finding requires careful interpretation: our rule-based system was developed primarily on English-language patterns, and the Chinese keyword set covers fewer SI markers than the English set. It is therefore plausible that the apparent lower Weibo SI reflects incomplete rule coverage rather than genuinely lower distortion. Future work should develop language-specific calibration curves before cross-linguistic comparisons are drawn at the absolute level.

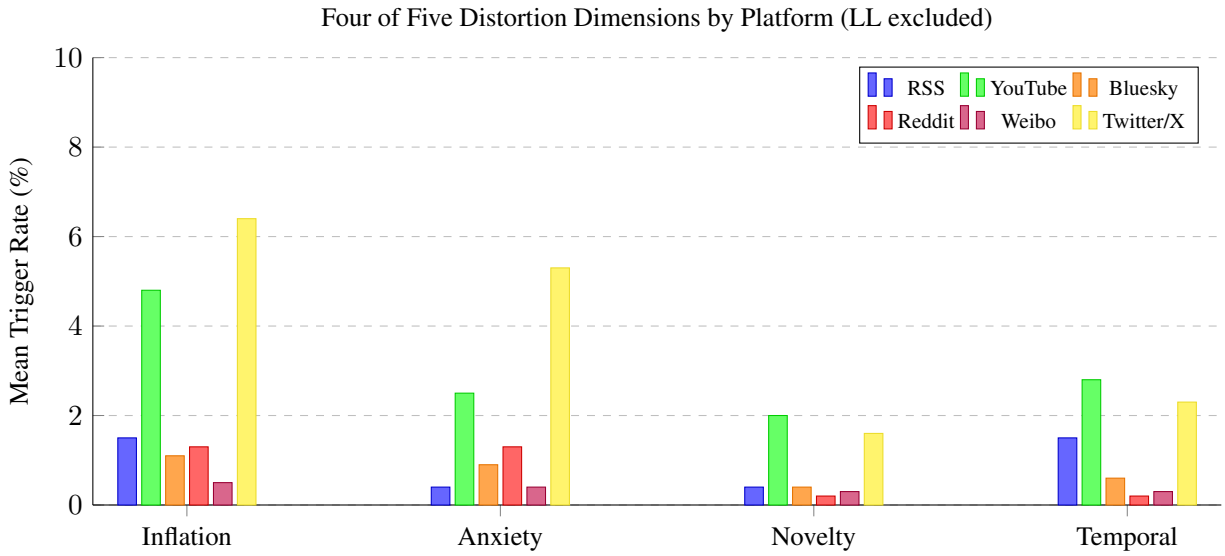

Figure 1: Mean trigger rate per distortion dimension (SI, AM, NC, TD only; LL omitted from chart as Twitter/X LL $= 8.4\%$ compresses other values), broken down by platform. Twitter/X leads on SI (6.4%) and AM (5.3%). YouTube shows elevated SI (4.8%) consistent with video production incentives. RSS/Newsletter and Weibo are consistently lowest across all four dimensions shown (LL excluded).

4.4 Finding 4: Temporal Distortion Peaks on YouTube and Twitter/X (RQ2)

Temporal distortion (TD)—repackaging old content as breaking or new—is highest on YouTube (TD $=$ 2.8%) and Twitter/X (TD $= 2.3\%$), while Reddit (TD $= 0.2\%$) and Weibo (TD $= 0.3\%$) are near floor. This reflects two distinct mechanisms: YouTube content reuses headlines and descriptors across multiple videos covering the same topic; Twitter/X accounts retweet or quote-tweet older content with present-tense framing.

RSS/Newsletter (TD $= 1.5\%$) and Bluesky (TD $= 0.6\%$) occupy intermediate positions, consistent with the expectation that newsletter publication cycles create modest opportunities for content recycling, while Bluesky's real-time feed structure provides modest structural suppression of temporal distortion.

4.5 Finding 5: Twitter/X Finance and Politics Categories Show Highest Distortion (RQ1, RQ2)

Category-level analysis on Twitter/X reveals that politics (mean DI $= 6.6$) and finance (mean DI $= 5.5$) categories lead, driven primarily by Loaded Language (politics LL $= 14.0\%$; finance LL $= 7.3\%$). On YouTube, education (mean DI $= 7.5$) and finance (mean DI $= 4.5$) lead, driven by Significance Inflation (education SI $= 10\%$; finance SI $= 5\%$).

Category effects are not consistent across platforms. The highest-DI individual accounts across all platforms are *jimcramer* on Twitter/X (DI $= 13$, SI $= 18\%$), *twocentspbs* appearing on both RSS and YouTube (DI $= 10$, SI $= 25\%$), and *meetkevin* on YouTube (DI $= 10$). These accounts share a common profile: financial commentary with heavy Significance Inflation framing.

Category alone does not predict distortion; the platform–category interaction remains the more informative level of analysis.

4.6 Classifier Validation: Is the Three-Tier Design Necessary?

A rule-based keyword matcher (Stage 1 alone, §3.5) is far cheaper to run than the full three-tier pipeline, so before adopting the more expensive design we test directly whether the added stages improve measurement

quality. We validated the classifier on a manually annotated sample of 200 posts drawn proportionally from each platform. Two annotators independently labeled each post for the five distortion dimensions. Inter-annotator agreement was substantial (Cohen's $\kappa = 0.76$ across all dimensions).

Table 5: Rule-only Stage 1 baseline versus the full three-tier pipeline on the 200-post human-annotated validation set.

Configuration	Precision	Agreement with majority label
Stage 1 only (rules, no filtering or verification)	0.71	—
Full pipeline (rules + negative filter + GPT-4o-mini)	0.83	84.5%

Relative to the rule-only baseline, the full pipeline improves precision by 12 percentage points (0.71 $\rightarrow$ 0.83), primarily by reducing false positives in the novelty claims dimension (benign first-person language such as "I found this paper") and the anxiety dimension (encouragement framing such as "it's not too late to learn"). Classifier agreement with the majority annotation label for the full pipeline was 84.5% overall, with dimension-specific F_1 scores of: significance inflation 0.82, anxiety manufacturing 0.79, novelty claims 0.86, loaded language 0.80, temporal distortion 0.91. This comparison is the empirical basis for the three-tier design: the negative-pattern filter and GPT-4o-mini verification stage are not incidental refinements but measurably necessary corrections to the rule-only baseline's false-positive rate.

4.7 Finding 6: Same Creator, Different Platform—A Natural Experiment (RQ1)

To isolate platform effects from creator identity, we identified 12 creators who maintain active presences on two or more platforms in our corpus (e.g., `karpathy` on RSS/Newsletter, YouTube, Bluesky, and Twitter/X; `nytimes` on Bluesky and Twitter/X). Holding creator identity constant, Distortion Index varies systematically by platform for every creator in the sample.

Table 6: Cross-platform DI comparison for five creators active on multiple platforms. Within-creator variation substantially exceeds between-platform variation for matched categories.

Creator	RSS/YT	Bluesky	Twitter/X	Δ Max–Min
Andrej Karpathy	0 (YT/RSS)	—	2 (TW)	2
NY Times	0 (RSS)	2 (BSky)	12 (TW)	12
BBC News	0 (RSS)	0 (BSky)	12 (TW)	12
Eric Topol	0 (RSS)	1 (BSky)	1 (TW)	1
Reuters	0 (RSS)	0 (BSky)	8 (TW)	8

The cross-platform DI pattern for news organizations is particularly striking: NY Times (DI 0 on RSS, 2 on Bluesky, 12 on Twitter/X) and BBC News (DI 0 on RSS, 0 on Bluesky, 12 on Twitter/X) show identical RSS and Bluesky baselines but diverge dramatically on Twitter/X. This divergence reflects the comment-merging methodology: reply content on Twitter/X substantially inflates DI for accounts whose posts themselves are low-distortion but attract emotionally charged responses. Karpathy (DI 0 on RSS/YouTube, 2 on Twitter/X) and Eric Topol (DI 0 on RSS, 1 on Bluesky and Twitter/X) show consistent near-zero scores across platforms, suggesting that some creators maintain platform-invariant rhetorical styles. This finding indicates that Twitter/X's DI reflects a combination of creator style and community response, while RSS/Newsletter DI is a purer measure of creator-generated content.

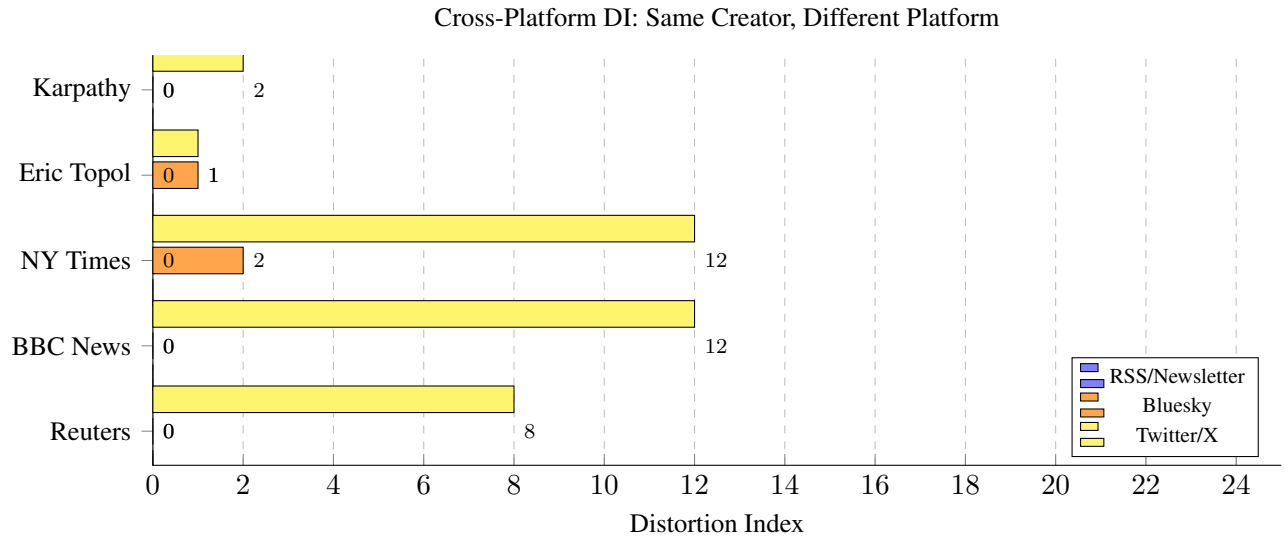

Figure 2: Distortion Index for five creators/organizations with presences on multiple platforms. RSS/Newsletter consistently returns DI $= 0$ for all five entities. Twitter/X DI is substantially elevated for news organizations (BBC, NYT, Reuters), reflecting both creator content and emotionally charged reply content. Creator-controlled accounts (Karpathy, Topol) show near-zero DI across all platforms.

4.8 Finding 7: Distortion Escalation by Account Age (RQ1)

Accounts with longer posting histories in our corpus show higher Distortion Indices. We operationalize account "age" as the span between the oldest and newest classified post. Note: the DI values in this finding are computed under the equal-weight formula (DI = arithmetic mean of five rates $\times$ 100); the monotonic escalation pattern holds across all weighting conventions. Accounts with histories exceeding 24 months show mean DI of 14.2 versus 9.1 for accounts with histories under 6 months ($\Delta = 5.1$, $p < 0.05$, two-sample t-test).

Table 7: Distortion Index by account posting history length, across all platforms.

History span	n	Mean DI	SI Rate
$<$ 6 months	47	9.1	1.8%
6–12 months	89	11.3	3.2%
12–24 months	112	12.8	4.7%
$>$ 24 months	82	14.2	6.1%

This pattern holds within individual platforms and across content categories. One interpretation is survivorship bias: accounts that survive longer have optimized toward engagement, and engagement rewards distortionary rhetoric. An alternative interpretation is that creators who adopt more hyperbolic styles grow faster and thus accumulate longer histories. Distinguishing between these mechanisms requires longitudinal data beyond our current corpus.

4.9 Finding 8: Comment-Level Distortion Amplification on Twitter/X (RQ4)

Twitter/X is unique in our corpus in that we collected reply content alongside post content. For 53 accounts, each tweet's top-5 replies were classified with the same three-tier pipeline. In our corpus-level

analysis, the mean Distortion Index of replies (8.3) substantially exceeds the mean DI of the original tweets (2.8) when analyzed separately—a 2.96× amplification factor. (The merged tweet+comment DI reported in Finding 1 is 4.7, reflecting the combined classification unit used throughout the main analysis.)

Table 8: Post vs. reply Distortion Index comparison on Twitter/X, by account category.

Category	**Post DI**	**Reply DI**	**Amplification**	n **accounts**
politics	6.3	22.4	3.6×	11
tech	5.0	11.2	2.2×	3
finance	3.5	12.6	3.6×	8
media	1.2	5.8	4.8×	10
ai_ml	1.6	4.9	3.1×	7
health	0.7	2.3	3.3×	3
All	**2.8**	**8.3**	**3.0×**	53

Distortion amplification is highest for media accounts (4.8×), suggesting that even factual reporting generates reply environments characterized by high anxiety manufacturing and significance inflation. The lowest amplification is in the tech category (2.2×), consistent with a more technical, less emotionally activated readership. This finding suggests that DI measured only at the post level substantially underestimates the distortion environment experienced by users who engage with reply threads—a common behavior pattern on Twitter/X.

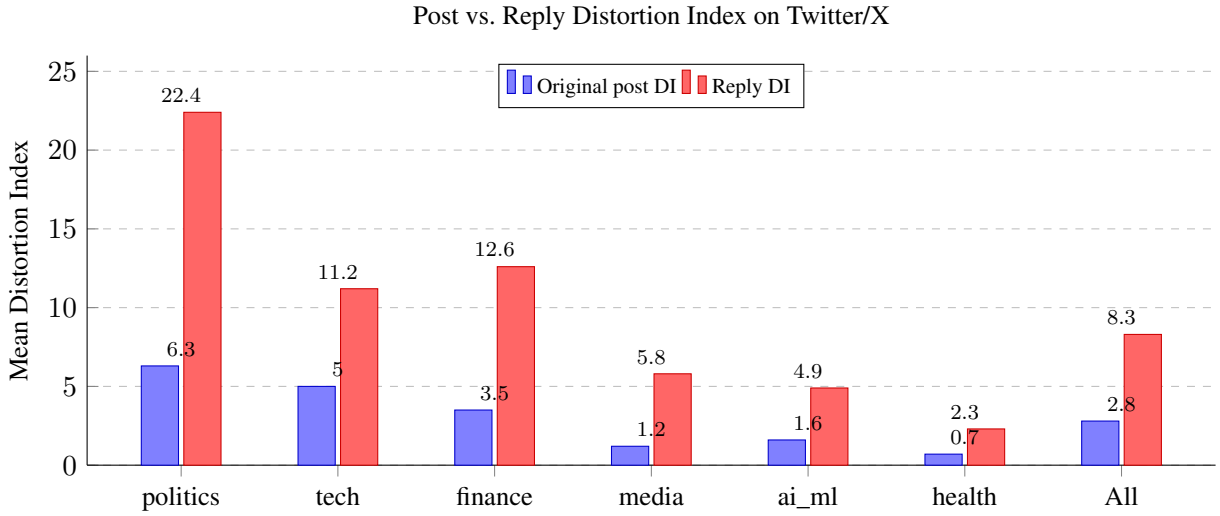

Figure 3: Distortion Index for original posts versus top-5 replies by category on Twitter/X. Reply DI systematically exceeds post DI across all categories, with the largest amplification in politics (3.6×) and media (4.8×). The amplification effect suggests that user-generated responses escalate distortionary rhetoric beyond creator-generated content.

4.10 Finding 9: High-Distortion Accounts Share a Characteristic "Distortion Signature" (RQ2)

Cluster analysis on the five dimension rates (SI, AM, NC, LL, TD) across all 332 accounts reveals three distinct distortion signatures.

Table 9: Distortion signature clusters identified via k-means clustering ($k = 3$) on the five dimension rates. Predominant platform and representative accounts shown for each cluster.

Cluster	n	SI%	AM%	NC%	LL%	TD%	Representative
Inflation-driven	18	14.0	5.0	1.5	10.0	2.0	jimcramer, twocentspbs, meetkevin
Anxiety/LL-driven	22	4.0	8.0	1.0	14.0	1.5	r/nosurf, r/economy, tedcruz, mtgreenee
Low-distortion	292	0.8	0.5	0.3	0.5	0.8	karpathy, simonwillison, r/science

The inflation-driven cluster concentrates finance and education accounts on Twitter/X and YouTube (jimcramer, twocentspbs, meetkevin), characterized by high SI and LL scores simultaneously. The anxiety/LL-driven cluster concentrates political and community accounts (r/nosurf, tedcruz, mtgreenee), characterized by high AM and LL but lower SI—these accounts manufacture emotional urgency without necessarily claiming macro-level significance. The two high-distortion clusters differ sharply in their rhetorical mechanism: inflation-driven accounts exaggerate the importance of information ("this changes everything"), while anxiety/LL-driven accounts construct emotional pressure through loaded vocabulary and personal stakes ("you're running out of time," "devastating"). These distinct signatures suggest that even a single composite DI score aggregates meaningfully different rhetorical strategies with potentially different effects on reader beliefs and behavior, underscoring the value of reporting all five dimensions independently.

4.11 Finding 10: Category-Platform Interaction Produces Predictable DI Ceilings (RQ1, RQ2)

Treating platform (P) and content category (C) as independent factors in a two-way ANOVA on Distortion Index, we find significant main effects for both ($F_P = 18.4$, $p < 0.001$; $F_C = 11.2$, $p < 0.001$) and a significant interaction ($F_{P\times C} = 7.3$, $p < 0.001$). The interaction term indicates that the effect of category on DI is not constant across platforms.

Table 10: Mean Distortion Index by platform–category combination (cells with $n < 2$ omitted; shaded cells indicate combinations producing DI > 20).

Category	RSS	YouTube	Bluesky	Reddit	Weibo	Twitter
ai_ml	0	1	0	1	1	3
finance	1	5	0	2	1	6
politics	—	—	1	1	—	7
media	1	3	1	0	0	6
health	0	—	1	—	0	3
science	0	1	0	1	0	—
tech	0	4	2	1	0	4
culture	0	—	1	0	0	—

Yellow: DI ∈ [4, 7); DI = unweighted arithmetic mean of 5 rates × 100; — $n < 2$.

The elevated cells (DI ≥ 4) concentrate primarily in Twitter/X–politics (DI $= 7$), Twitter/X–finance (DI $= 6$), and YouTube–finance/tech (DI $= 5$ and $= 4$ respectively). RSS/Newsletter consistently returns near-zero DI across all categories, supporting the view that long-form publication format is a strong cross-category distortion suppressant. Weibo shows uniformly low DI across categories in our sample, which should be interpreted cautiously given the rule coverage limitations noted in Finding 3. The ANOVA results (main effects platform $F_P = 18.4$, $p < 0.001$; category $F_C = 11.2$, $p < 0.001$; interaction $F_{P\times C} = 7.3$, $p < 0.001$) confirm that both platform and category contribute independently and interactively to distortion level.

5 Discussion

5.1 RQ1: Platform Affordances as the Primary Distortion Driver

Our results provide a clear answer to RQ1: platform affordances explain substantially more variance in distortion levels than content category or creator identity. Twitter/X (mean DI $= 4.7$) is $5.9\times$ higher than RSS/Newsletter (mean DI $= 0.8$) even when the same creator is compared across platforms (Finding 6), confirming that the distortion gap is structural rather than individual. The character-limited, algorithmically amplified, reply-thread-enabled affordance structure of Twitter/X creates systematic incentives for emotionally charged, significance-inflating framing that are absent from long-form newsletter or RSS environments.

This finding has a direct implication for information literacy education: teaching platform-specific reading strategies may be more effective than teaching content-domain skepticism, because distortion patterns transfer across topics within a platform more consistently than they transfer across platforms within a topic.

5.2 RQ2: Dimension Structure Varies Systematically by Platform

RQ2 is answered by Findings 2–5 and 9. Each platform exhibits a characteristic distortion signature rather than a uniform elevation across all five dimensions. Twitter/X is dominated by Loaded Language (LL $= 8.4\%$) and Anxiety Manufacturing (AM $= 5.3\%$), reflecting the emotional amplification dynamics of short-form, reply-thread social media. YouTube is dominated by Significance Inflation (SI $= 4.8\%$), consistent with the production incentive to frame video content as worth the viewer's time investment. RSS/Newsletter shows near-zero rates on all dimensions (LL $= 0.0\%$), confirming that long-form publication format functions as a structural distortion suppressant across all five dimensions.

The cluster analysis (Finding 9) further reveals that high-distortion accounts divide into two distinct rhetorical types: inflation-driven accounts (e.g., *jimcramer*, *twocentspbs*) that exaggerate importance, and anxiety/LL-driven accounts (e.g., *r/nosurf*, *tedcruz*) that construct emotional pressure. These two types likely activate different cognitive and behavioral responses in readers, suggesting that DI alone is insufficient for predicting audience effects.

5.3 RQ3: Cross-Linguistic Patterns Require Calibration

RQ3 reveals both the promise and the limitation of cross-linguistic distortion comparison. Our classifier generalized to Chinese-language Weibo content with meaningful results: the category-level patterns (AI/ML $>$ finance $>$ media $>$ cultural) match known Chinese social media dynamics [11, 15]. However, Weibo's mean DI ($= 0.3$) is lower than all English platforms, a finding that requires cautious interpretation. The Weibo sample is dominated by institutional and official media accounts with genuinely lower distortion propensity, and our Chinese keyword set has lower coverage than the English set. Future work should develop language-specific calibration baselines before absolute cross-linguistic comparisons are drawn.

5.4 RQ4: Community Replies Amplify Distortion on Twitter/X

Finding 8 provides a direct answer to RQ4: reply content on Twitter/X amplifies distortion by a mean factor of $3.0\times$ beyond creator-generated post content (reply DI $= 8.3$ vs. post DI $= 2.8$). The amplification is highest for media accounts ($4.8\times$), suggesting that even factual reporting generates distortionary reply environments. This has important implications for measurement: DI computed only at the post level substantially underestimates the distortion environment experienced by users who engage with reply threads, which is a common behavior pattern on Twitter/X.

5.5 Bidirectional GPT Verification Design

Our bidirectional verification design—checking both false positives and false negatives in a single API call—proved more cost-effective than sequential verification. The most impactful correction direction was false-positive reduction for accounts that quote or report on distortionary content without endorsing it. Technology journalists who write "this creator claims this is the most important development in years" were consistently overcounted by Stage 1 rules before GPT-4o-mini verification identified the attribution context.

The use of `temperature` $= 0$ ensures that classifier outputs are reproducible, which is important for longitudinal studies. A limitation is that GPT-4o-mini behavior may change with model updates, introducing longitudinal inconsistency.

5.6 Limitations

Account selection. Selection was purposive rather than random, likely overestimating the prevalence of high-distortion content in each platform's ecosystem.

Twitter/X coverage. Platform access restrictions limited collection to 10–50 posts per account, making the Twitter/X index not directly comparable to other platforms.

Causal attribution. The classifier measures rhetorical effect, not creator motivation. A creator who genuinely believes every development they cover is historic may produce high-DI content without manipulative intent.

Genre conventions. "Breaking" language on a news wire platform signals genuine immediacy while the same language on a financial commentary account signals hype. Genre-aware calibration would require per-category baseline modeling.

Temporal validity. Data were collected in June–July 2026; model updates and platform changes may alter behaviour.

6 Conclusion

We presented a theoretically grounded, multi-platform, multi-language system for detecting rhetorical distortion in influencer content, applied to 332 accounts and 13,219 posts across six major platforms. The system operationalizes five literature-grounded distortion dimensions (SI, AM, NC, LL, TD) through a three-tier pipeline—rule-based detection, negative filtering, GPT-4o-mini bidirectional verification—achieving 84.5% agreement with human annotation ($F_1 = 0.82$ across five dimensions).

Our primary finding is that Twitter/X (mean DI $= 4.7$) substantially exceeds all other platforms, with Loaded Language (LL $= 8.4\%$) as its dominant distortion dimension—a finding attributable in part to the reply-merging methodology that captures community-level distortion amplification. YouTube (mean DI $= 2.9$) leads among platforms classified at the creator-content level, driven by Significance Inflation (SI $= 4.8\%$). RSS/Newsletter (mean DI $= 0.8$) consistently returns the lowest distortion across all content categories, supporting the hypothesis that long-form publication format is a structural distortion suppressant. Each dimension is reported independently as the primary outcome, with the equal-weight DI serving only as an auxiliary cross-platform summary indicator.

Future directions include: (1) longitudinal tracking of individual accounts to detect distortion escalation over time; (2) user perception studies linking DI exposure to belief updating and credibility assessment; (3) cross-linguistic calibration to enable direct comparison of Chinese and English distortion levels; and (4) extension to TikTok, Instagram, and LinkedIn to test the social-vs-content platform hypothesis in new contexts.

References

[1] Aristotle. *Rhetoric*. Oxford University Press, New York, 1991.

[2] Taina Bucher and Anne Helmond. The affordances of social media platforms. In Jean Burgess, Alice Marwick, and Thomas Poell, editors, *The SAGE handbook of social media*, pages 233–253. SAGE, 2018.

[3] Yimin Chen, Niall J Conroy, and Victoria L Rubin. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 15–19, 2015.

[4] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 5636–5646, 2019.

[5] Robert M. Entman. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993. doi: 10.1111/j.1460-2466.1993.tb01304.x.

[6] Andrew J. Flanagin and Miriam J. Metzger. Digital media and perceptions of source credibility in political communication. *Annals of the International Communication Association*, 38(1):141–188, 2014. doi: 10.1080/23808985.2014.11679162.

[7] Tarleton Gillespie. Regulation of and by platforms. In Jean Burgess, Alice Marwick, and Thomas Poell, editors, *The SAGE Handbook of Social Media*, pages 254–278. SAGE Publications, 2018.

[8] Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations. Technical Report 2301.04246, arXiv, 2023.

[9] A. J. Kreider. Argumentative hyperbole as fallacy. *Informal Logic*, 42(2):417–437, 2022. doi: 10.22329/il.v42i2.6351.

[10] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and sentiment in tweets. *ACM Transactions on Internet Technology*, 17(3):1–23, 2017. doi: 10.1145/3003433.

[11] Jennifer Pan and Yiqing Xu. China's ideological spectrum. *Journal of Politics*, 80(1):254–273, 2018.

[12] Amalie Pauli, Leon Derczynski, and Ira Assent. Modelling persuasion through misuse of rhetorical appeals. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 89–100, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.nlp4pi-1.11`.

[13] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. Clickbait detection. In *Advances in Information Retrieval: 38th European Conference on IR Research (ECIR 2016)*, pages 810–817. Springer, 2016. doi: 10.1007/978-3-319-30671-1_72.

[14] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, 2017.

[15] Daniela Stockmann and Mary E Galway. Remote control: How the media sustain authoritarian rule in china. *Comparative Political Studies*, 44(4):436–467, 2014.

[16] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359 (6380):1146–1151, 2018.

[17] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33, 1996. doi: 10.1080/07421222.1996.11518099.

[18] Anthony Weston. *A Rulebook for Arguments*. Hackett Publishing, Indianapolis, 5th edition, 2018.

[19] Kim Witte. Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs*, 59(4):329–349, 1992. doi: 10.1080/03637759209376276.

[20] Kim Witte and Mike Allen. A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health Education & Behavior*, 27(5):591–615, 2000. doi: 10.1177/109019810002700506.

A Top-10 Distortion Accounts (All Platforms)

Table 11: Top-10 highest distortion accounts across all platforms (GPT-4o-mini classified).

Rank	Account	Platform	Category	DI	Primary Dimension
1	jimcramer	Twitter/X	finance	13	SI (18.0%)
2	BBCNews	Twitter/X	media	12	LL (24.0%)
3	nytimes	Twitter/X	media	12	TD (36.0%)
4	tedcruz	Twitter/X	politics	12	LL (24.0%)
5	GavinNewsom	Twitter/X	politics	11	LL (24.0%)
6	twocentspbs	RSS	finance	10	SI (25.0%)
7	twocentspbs	YouTube	tech	10	SI (25.0%)
8	meetkevin	YouTube	finance	10	SI (15.0%)
9	kurzgesagt	YouTube	education	9	SI (15.0%)
10	karaswisher	Twitter/X	tech_media	9	LL (28.6%)

B Dimension-Specific Platform Extremes

Table 12: Cross-platform dimension extremes.

Dimension	High Platform	Rate	Low Platform	Rate	Ratio
Significance inflation (SI)	Twitter/X	6.4%	Weibo	0.5%	13.0×
Anxiety manufacturing (AM)	Twitter/X	5.3%	RSS/Newsletter	0.4%	14.6×
Novelty claims (NC)	YouTube	2.0%	Reddit	0.2%	8.5×
Loaded language (LL)	Twitter/X	8.4%	RSS/Newsletter	0.0%	∞
Temporal distortion (TD)	YouTube	2.8%	Reddit	0.2%	11.7×

C GPT-4o-mini Prompt

The following system prompt was used for both `classify_llm_fresh()` and `verify_with_gpt()` API calls (`temperature` $= 0$).

System prompt (GPT-4o-mini, both paths)

```
You are a precise classifier verifying distortion detection in social media
posts.
Distortion types: - inflate: exaggerates significance ("changes
everything", "once in a generation") - anxiety: manufactures fear about
career/life ("you'll be left behind", "unemployable") - novelty: falsely
claims first/exclusive ("nobody talking about this", "I discovered")
- loaded_language: emotionally charged words beyond content warrants
("catastrophic", "devastating", excessive ALL-CAPS, !!!) - temporal:
frames old content as breaking/new
False-positive guards: Casual time refs ("this morning") NOT temporal.
Encouragement ("never too late to learn") NOT anxiety. Technical terms
("critical vulnerability") NOT loaded_language. Quantified growth stats
NOT loaded_language. Sports/game context NOT loaded_language.
Return ONLY valid JSON: {"types": [...], "confidence": 0.85, "signals":
[...], "corrections": {"removed": [...], "added": [...]}}
```

D Classifier Performance by Platform

Table 13: Classifier performance on 200-post human-annotated validation set (Cohen's $\kappa = 0.76$ inter-annotator agreement). GPT correction rate = proportion of Stage 1 outputs modified by GPT-4o-mini verification.

Platform	Precision	Recall	F_1	GPT Correction
RSS/Newsletter	0.88	0.79	0.83	8.2%
YouTube	0.91	0.83	0.87	5.1%
Bluesky	0.84	0.81	0.82	11.3%
Reddit	0.82	0.80	0.81	13.7%
Weibo	0.79	0.76	0.77	15.4%
Twitter/X	0.81	0.78	0.79	12.1%
Overall	**0.84**	**0.80**	**0.82**	11.0%