Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Collective belief revision as a saturating function of accumulated knowledge: cross-platform evidence from social media

Yuexin Lan
China People's Police University
Bo Mao
China People's Police University
Sijia Li
China People's Police University
Qifeng Li
Labc11@126.com

Guangdong Police College
Shuangshi Zhang
China People's Police University

Article

Keywords: collective belief, knowledge accumulation, belief revision dynamics, misinformation correction, stance analysis, threshold models

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11420824/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Collective belief revision as a saturating function of accumulated knowledge: cross-platform evidence from social media

Abstract

Probability is usually treated as static given the knowledge that conditions it. This paper proposes a dynamical interpretation of collective belief — expressed belief as a saturating function of accumulated knowledge V with a revision ceiling K_m (working sigmoid form $p = K_m \cdot \sigma(\alpha V + \beta)$) — tested on 297 Twitter rumour threads (PHEME/RumourEval-2017) and 3,387 Sina Weibo repost cascades (CED). Pooled revision curves are saturating and monotone; they survive within-thread, cross-thread and block-permutation placebos ($p \leq 0.02$), de-cumulation, partial-pooling and leave-one-event-out refits. The ceiling separates false from true rumours by ≈2-fold (1.7–2.4 under sensitivity analyses; paired event-cluster bootstrap P(reversal) = 0.012; with 7/5 usable events per side, event-level inference remains resolution-limited by design); non-rumour cascades are flat. Because comment streams under true claims are not truth-conducive, the contrast is read as a correctability measure for false claims only; on Weibo the ceiling is measurement-conditioned ($K_m \approx 0.50$–0.92 across denominator corrections). The logistic family is not identified, so only family-independent findings are claimed: a ceiling, monotone saturation, and sequence-dependence.

Keywords: collective belief; knowledge accumulation; belief revision dynamics; misinformation correction; stance analysis; threshold models

1. Introduction

Whether probability describes the world or our knowledge of the world is one of the oldest disputes in the foundations of inference. The epistemic tradition — from Laplace's demon through Keynes (1921), Cox (1946), de Finetti (1974), and Jaynes (2003) — holds that probability is a degree of belief warranted by available information: were knowledge

complete, probability would collapse into certainty. Yet this tradition has remained largely static. Probability theory tells us how beliefs must cohere at a fixed state of information; it says little about how belief evolves as knowledge accumulates, and nothing about the shape of that evolution.

This paper investigates a dynamical completion of the epistemic view: collective belief is a saturating function of accumulated knowledge, with working sigmoid form $p = K_m \cdot \sigma(\alpha V + \beta)$, where V measures the knowledge a community has accumulated about a claim, p its belief state (the share of expressed stances that deny the claim), and K_m an intrinsic revision ceiling. Three literatures touch pieces of this idea without assembling it. Psychometrics maps a latent trait onto response probabilities through a logistic link (Rasch, 1960); random-utility models aggregate heterogeneous thresholds into logistic choice shares (McFadden, 1974); diffusion research finds S-shaped adoption (Rogers, 1962; Bass, 1969); Granovetter showed that a population's adoption curve is the distribution function of individual thresholds (Granovetter, 1978), and Watts derived how threshold distributions gate cascades (Watts, 2002). Collective-intelligence and prediction-market work models aggregated judgement as a function of pooled information (Hong and Page, 2004; Wolfers and Zitzewitz, 2004), and belief-as-accumulated-log-likelihood-ratio with constant drift is standard in decision neuroscience (Gold and Shadlen, 2007; Glaze et al., 2015) — the content of our bridging assumption. What is missing is the unifying claim — that these are shadows of one object, a belief measure that is the integral of knowledge accumulation — with its falsifiable consequences.

We position our contributions narrowly. The mathematical core (§2.1, Propositions 1–3) restates known results — the odds form of Bayes' theorem (Good, 1950; Jaynes, 2003), the equal-covariance discriminant identity, and the threshold/distribution-function correspondence underlying item response theory, random-utility models, and sociological

threshold models (Rasch, 1960; McFadden, 1974; Granovetter, 1978) — because the synthesis is part of the message: read together, these results say the sigmoid map is not a metaphor but the generic form of aggregated Bayesian revision. The architecture is not unprecedented either: MacCoun's BOP/BEAN models already combine a logistic threshold form, an evidence axis, and a ceiling parameter structurally analogous to K_m (MacCoun, 2012, 2015). Two ingredients are, to our knowledge, not explicit elsewhere. First, Proposition 4 collects elementary differential properties of the logistic curve — two parameter-free transitions at $\frac{\left(1\pm\frac{1}{\sqrt{3}}\right)}{2}$ of the ceiling, with the central regime carrying exactly $\frac{1}{\sqrt{3}} \approx 57.7\%$ of the total increment — and repurposes them as a family-discrimination device, since the corresponding probit thresholds (0.159/0.841) differ materially. The properties are standard in the growth-curve literature (Richards, 1959); the test design built on them is new. Second, we introduce a cross-platform, veracity-stratified ceiling-estimation design: the same scalar K_m estimated under veracity labels on Twitter and against a non-rumour control on Weibo, under a battery of measurement and inference stress tests. The bridging assumption is stated as Axiom A1: each unit of accumulated knowledge carries, in expectation, a constant amount of log-evidence — a first-moment statement, not an identity.

A second anchor for the revision ceiling $K_m < 1$ comes from social learning and opinion dynamics. Information-cascade models show that rational agents can collectively fail to learn (Bikhchandani et al., 1992; Banerjee, 1992); Smith and Sørensen prove that even with unbounded signals a Bayesian population converges to the wrong action with positive probability (Smith and Sørensen, 2000); Friedkin–Johnsen dynamics endow agents with a stubbornness parameter near-isomorphic to our never-converters (Friedkin and Johnsen, 1990, 1999); and the SEIZ model's Skeptic compartment is the epidemic-side precedent of a sub-one ceiling (Jin et al., 2013). Our framework makes the matching micro-move explicit: if

individual thresholds follow a defective distribution — a point mass $1 - K_m$ at $+\infty$ — the population curve is exactly $p(V) = K_m \cdot F(V)$, and incomplete correction becomes a parameter rather than an anomaly (§2.1, Proposition 3). We stress that this reading is one of several nearly exchangeable mechanisms, not an identification result.

On the empirical side we bring two independent datasets — 297 Twitter rumour threads with 4,519 crowd-expert stance annotations (PHEME/RumourEval-2017) and 3,387 Sina Weibo repost cascades with 1.28 million reposts (CED) — and show that pooled collective belief revision follows saturating, monotone curves. The fitted ceiling discriminates false from true information by a factor of ≈2 on Twitter (0.37 vs 0.17; 1.7–2.4 across inclusion thresholds and leave-one-event-out refits), with a paired event-cluster bootstrap giving $P(K_m_false \leq K_m_true) = 0.012$ (B = 500); on Weibo, non-rumour cascades show no revision dynamics. The curves are genuinely sequential (within-thread stance-permutation placebos, $p \leq 0.02$ across variants; $p \leq 0.009$ for the primary volume-based analyses) and are not an artefact of cumulative averaging. We are deliberately more cautious about everything family-dependent: pooled in-sample comparisons lean logistic over probit on the large Weibo sample, but non-sigmoid saturating baselines tie the logistic on both Weibo curves, out-of-sample prediction cannot separate the two symmetric families, and a Richards asymmetry fit pins at its boundary — so we report the family question as unresolved, as Chambers and Cox (1967) warn that discrimination needs tail information. We also report, without cosmetic repair, that the latent regime predicted by the logistic anatomy is not observed on either platform. Both shape and ceiling conclusions survive two measurement upgrades — a human-validated supervised stance filter for the Chinese data and a content-based knowledge axis — plus the robustness battery of §2.5.

The remainder of the paper states the framework and propositions (§2.1), reports the empirical results (§2.2–2.6), and discusses implications, boundary conditions, and limitations (§3). Methods and robustness details follow (§4).

2. Results

Table 1 summarises the two datasets analysed in §2.2–2.6: PHEME/Twitter rumour threads and CED/Sina Weibo repost cascades.

Table 1. Dataset summary.

Property	PHEME / RumourEval-2017 (Twitter)	CED (Sina Weibo)
Unit of analysis	rumour thread	repost cascade
Threads / cascades	297 (train+dev release of SemEval-2017 Task 8)	3,387 (1,538 rumour + 1,849 non-rumour)
Posts	4,519 tweets	1,275,106 reposts (768,551 with text; 506,555 text-less; authors' count, §4.1)
Period	2014–2015	2009–2015
Events	8 breaking-news events (Table 3)	—
Stance labels	human annotation (support/deny/query/comment)	two-stage: lexicon + supervised filter (§4.3–4.4)
Veracity labels	thread-level (137 true / 62 false / 98 unverified)	collection design (rumour vs non-rumour; platform-adjudicated, §4.1)
Inclusion (primary)	≥ 3 stance posts → 184 threads (false 39 / true 80 / unverified 65)	≥ 15 stance posts → rumour 245 (classifier) / 329 (lexicon), non-rumour 30 / 39

2.1 A dynamical interpretation: probability as a saturating map of accumulated knowledge

We consider a community confronting a binary claim (e.g., "this rumour is true" vs "false"). Let V denote accumulated knowledge — operationally, the cumulative, de-duplicated information bearing on it (normalised information progress within a thread; content-based units in Supplementary Note S5) — and p(V) the collective belief state, the share of expressed stances that deny the rumour. The framework asserts $p(V) = K_m \cdot \sigma(\alpha V + \beta)$, equivalently $K_m \cdot \sigma\big(\alpha(V - V_0)\big)$ with $\beta = -\alpha V_0$; we use the V_0 parameterisation throughout. One axiom and four propositions (proofs in S1) fix its content: A1 bridges theory

and measurement; Propositions 1–3 restate classical results because their conjunction fixes what accumulates and what the link identifies; Proposition 4 repurposes elementary properties as a family-discrimination device.

Sign and direction: what α measures. α is a signed empirical quantity: it records both the strength and the direction of the evidence stream ($\alpha < 0$ for truth-conducive streams bearing on false claims), and its units are log-odds per unit V. Its absolute value is not identified from shape alone; the exponent law of Proposition 4(iv) and segmented estimation (§2.6) supply the sharp tests.

Axiom A1 (constant expected log-evidence — the bridge). Each unit of accumulated knowledge carries, in expectation, a constant amount α of log-evidence: $E[\lambda|\Delta V] = \alpha \cdot \Delta V$. (i) It is a first-moment condition only. With the realised path $L(V) = \beta + \alpha V + M(V)$, M a zero-mean martingale of variance $\tau^2 V$, convexity below the inflection and concavity above give $E\left[\sigma\left(L(V)\right)\right] \neq \sigma(\alpha V + \beta)$ with a sign-changing bias: the aggregated curve sits above the deterministic curve early and below it late, so A1 licenses "log-odds accumulate linearly in expectation" and nothing stronger. (ii) If V is defined as the stock of log-evidence, A1 is vacuous; for a measured proxy it holds only in expectation. (iii) On the normalised axis any smooth monotone drift is absorbed by the logistic's three free parameters (placebo nulls reach $R^2 \approx 0.9$ under chance), so pooled fit is weak evidence; the sharp tests are the $\gamma = 1$ exponent law and segmented estimation (§2.6). All Bayesian readings are conditional on A1.

Proposition 1 (odds-form Bayes — classical). For a Bayesian agent with prior odds in $(0, \infty)$, posterior belief is the sigmoid of accumulated log-evidence: $p = \sigma(V)$ with $V = \log[\text{prior odds}] + \Sigma_i \lambda_i$, where each λ_i is the log-likelihood ratio of the i-th item (Good, 1950; Jaynes, 2003). Under A1, logit(p) is the stock of evidence and p its bounded image; the K_m =

1 case saturates at certainty as V → ∞. The prior-odds premise connects to Proposition 3d: a point mass at V* = +∞ is the population image of a dogmatic-prior subpopulation.

Proposition 2 (Gaussian bridge — classical). Equal-covariance Gaussian features yield a logit that is affine in the discriminant; statement and proof in Supplementary Note S1.

Proposition 3 (threshold micro-foundation — classical). Suppose each agent converts when accumulated knowledge exceeds an idiosyncratic threshold V*_i. The population share of converts is $p(V) = P(V^* \leq V) = F(V)$: the collective curve is the threshold distribution function — the correspondence underlying Rasch models, random-utility theory, and Granovetter's threshold model (Rasch, 1960; McFadden, 1974; Granovetter, 1978) — and the link function identifies the family (logistic ↔ logit; normal ↔ probit) up to location–scale equivalence, for a single population only. Three corollaries follow. (3b) Link comparison tests threshold families, though Chambers and Cox warn that discrimination is tail-driven and needs enormous samples (Chambers and Cox, 1967). (3c) The sigmoid is a population-level object; individual threads are noisy draws (§2.5). (3d) If thresholds are defective — a point mass $1 - K_m$ at V* = +∞ — then $p(V) = K_m \cdot F(V)$ and the curve saturates strictly below 1. This is one of three nearly exchangeable mechanisms (small α, Jensen attenuation), not an identification result; the point mass equals a dogmatic-prior subpopulation (Lewandowsky et al., 2012; Ecker et al., 2022), and a plateau below 1 with evidence still flowing would exclude the small-α alternative in absolute units.

Normalisation and pooling. Our measurement normalises V to [0,1] within each thread ($V_k = \frac{k}{n}$). With A1 this implies every thread carries the same total log-evidence regardless of length — a strong, previously unstated assumption; length then modulates only the resolution at which the curve is traced. If exchange rates or locations vary across threads, the pooled curve is a mixture of sigmoids and generally not logistic: pooled comparison identifies the

convolution of threshold and heterogeneity distributions, so Proposition 3b's single-population premise fails and §2.4 carries a homogeneity caveat. Simulations in §2.5 and S8 show that heterogeneity alone can generate high-R^2 pooled sigmoids.

Proposition 4 (intrinsic three-segment structure, only partially observable — elementary; the test design is new). For $p(V) = K_m \cdot \sigma(\alpha V + \beta)$ with α > 0: (i) $\frac{d^3p}{dV^3} = 0$ at $p^* = K_m \frac{\left(1 \pm \frac{1}{\sqrt{3}}\right)}{2} \approx 0.211 K_m$ and $0.789K_m$ — fractions of the estimated ceiling, independent of all parameters; (ii) the three segments carry belief increments in proportions $\frac{\left(1-\frac{1}{\sqrt{3}}\right)}{2} : \frac{1}{\sqrt{3}} : \frac{\left(1-\frac{1}{\sqrt{3}}\right)}{2}$, so the central segment always carries $\frac{1}{\sqrt{3}} \approx 57.7\%$ of total revision; (iii) the central segment spans $\Delta V = \frac{2\ln(2+\sqrt{3})}{\alpha} \approx \frac{2.634}{\alpha}$ in knowledge units, its width governed solely by α; (iv) the sign pattern of $\frac{d^3p}{dV^3}$ (+/−/+) marks the boundaries as inflection points of the impact curve $\frac{dp}{dV} = \left(\frac{\alpha}{K_m}\right) p(K_m - p)$, which peaks at $p = \frac{K_m}{2}$ — new knowledge strikes hardest where belief is maximally unsettled; (v) if $p_0 \geq 0.211K_m$ the system enters observation inside the high-response regime. Properties (i)–(v) are standard differential properties of the logistic (Richards, 1959); their use as an anatomical signature is the contribution, deliberately limited: the probit thresholds ($0.159K_m$, $0.841K_m$) differ by about five points, so the families are in principle distinguishable given sufficiently informative data — whether ours suffice is answered cautiously in §2.4. Two caveats frame what follows. First, the thresholds are specific to the symmetric logistic; a Richards (asymmetric) sigmoid moves them by several points (Richards, 1959). Second, both datasets enter observation late (Weibo $p_0/K_m \approx 0.71$–0.76; Twitter ≈ 0.83 after the small-denominator correction, §2.2), so the latent regime is never observed and we describe the anatomy as extrapolated.

Figure 1 summarises the framework: data accumulate into information, information into knowledge V; belief p is the sigmoid image of V; two parameter-free thresholds partition its anatomy.

2.2 Collective belief revision is saturating, monotone, and sequence-dependent

We first ask whether collective belief revision has the predicted sigmoid shape. For each thread (Twitter) or cascade (Weibo), we order stance-bearing posts by time, define V as the normalised position within the thread's information stream, and track the cumulative denial share p. Because individual threads are short and noisy (Proposition 3c), our primary analysis pools threads: each curve is interpolated onto a common 50-point grid and averaged; uncertainty comes from resampling threads (bootstrap, BCa for PHEME) and, because Twitter threads nest within 8 events, from resampling events and leave-one-event-out refits (§4.6).

On Twitter (PHEME; false rumours, n = 39), the pooled curve rises from ≈ 0.10 and saturates: the three-parameter logistic gives $K_m = 0.372$ (BCa 95% CI [0.282, 0.478]; event-cluster [0.253, 0.499]), inflection $V_0 = 0.245$, $R^2 = 0.984$ (Fig. 2a). The inflection is interior in a thread-level bootstrap (median 0.258, 95% CI [0.078, 0.468]; 98.7% of replicates inside the observed range), although the raw start falls below the lower transition in only 34.7% of replicates and is artefact-depressed ($p_0/K_m \approx 0.83$ after correction). On Weibo (CED), the classifier-filtered measurement (n = 245) yields $K_m = 0.502$ [0.467, 0.690], $R^2 = 0.994$; the lexicon measurement gives the same shape (n = 329, $K_m = 0.619$ [0.580, 0.949], $R^2 = 0.992$) (Fig. 2b). Because the support side is contaminated (audit precision 0.10), both are co-primary: correcting the denominator raises the Weibo ceiling to 0.823–0.919, with the strict-support variant at 0.851 (Table 2). Two platforms, languages, and sampling mechanisms thus produce the same qualitative object — a saturating, monotone function of accumulated

information, not a proportional response — though the lower half of the S is unobserved on either platform (§2.4).

We are explicit about what pooled R^2 does and does not establish. The grid points are interpolated, autocorrelated pseudo-observations and cumulative shares are smooth by construction; under a pure-chance null the same pipeline yields R^2 as high as ≈ 0.9. We therefore never rest a claim on R^2 alone: the evidence lies in the observed–null gap (§2.5), the flat non-rumour control (§2.3), shape-specific placebo statistics (§2.4–2.5), and de-cumulated shares (§2.5).

The Weibo curve starts well above zero (initial denial share 0.36–0.47 depending on measurement): these cascades originate from a rumour-reporting platform, so early posts already carry scepticism. Quantitatively, p_0/K_m = 0.76 (lexicon) and 0.71 (classifier), just below the upper transition 0.789 — Weibo cascades enter near the top of the high-response regime, and we observe essentially their upper segment only. The Twitter curve, rising from ≈ 0.10, looks closer to a full S, but two arithmetic checks discipline that reading (S8). First, the pooled start is mechanically depressed by small denominators: dropping each thread's first three stance observations raises it from 0.103 to 0.529 (Weibo moves only 0.357 → 0.374). Second, the first 30%-window mean of instantaneous denial share is already 0.31. The three entry values measure different quantities; the first-window share (0.31) is least biased. We therefore take $p_0 \approx 0.31$, i.e. $p_0/K_m \approx 0.83$, at or just above the upper transition $0.789K_m \approx 0.29$. On Twitter as on Weibo, observation begins late; the latent and most of the high-response regime are extrapolated, and the apparent acceleration is largely the small-denominator footprint. Two consequences follow: the lower non-parametric change point of §2.4 ($0.299K_m$) is likely contaminated; and the fitted ceiling is entangled with the artefact — the windowed share ends at ≈ 0.46, above $K_m = 0.372$ — so the Twitter ceiling is best read as

a lower bound. The early-entry corollary is a post-hoc consistency check, not an independent prediction.

2.3 The revision ceiling discriminates false from true information

If the ceiling K_m measures how far collective cognition can be corrected, it should differ by veracity. On Twitter it does: $K_m = 0.372$ for false rumours versus 0.172 [0.128, 0.227] for true ones, with unverified claims at 0.314 [0.253, 0.378] — closer to the false than to the true ceiling. The contrast survives conservative event-level inference: the event-cluster bootstrap (7 of 8 false-side events, 5 of 8 true-side; B = 500) gives CIs of [0.253, 0.499] and [0.131, 0.212], and the paired event-cluster bootstrap gives $P(K_m_false \leq K_m_true) = 0.012$ (k = 5 of 500), with Δ median 0.208 and 95% interval [0.069, 0.311]. Small-cluster percentile intervals under-cover (Cameron et al., 2008), so we complement both with two event-exact analyses (Fig. 3, Table 3).

First, leave-one-event-out refits: dropping any single event keeps the false/true ratio within [1.98, 2.41] (false K_m 0.331–0.434; true 0.164–0.183), so no single event dominates. Second, an exact sign-flip randomisation test at the event level: only G = 4 events have ≥ 3 usable threads on both sides, with within-event differences $d_e = -0.009$ (charliehebdo), +0.353 (germanwings-crash), +0.318 (ottawashooting), +0.327 (sydneysiege), giving two-sided p = 0.25. With G = 4 the minimum achievable p is 0.125, so event-level exact inference cannot reach significance with present data; the LOEO range plus 3 of 4 events pointing the same way is the strongest evidence form available. The charliehebdo anomaly is itself informative: it is the largest event, and within it the false ceiling (0.162) sits at the true ceiling (0.171) — the one event in which falsehoods were no more correctable than truths (§3).

The contrast is not an artefact of the knowledge proxy: under a content-based axis counting only semantically de-duplicated units (§4.5, S5), the ceilings replicate almost

exactly — false 0.390 [0.305, 0.495], unverified 0.308 [0.256, 0.365], true 0.181 [0.133, 0.235] — preserving the ≈2.2-fold ratio (Fig. 3). On Weibo, non-rumour controls show no revision dynamics: their pooled curves are flat under every measurement variant and inclusion threshold (total excursion ≤ 0.07 versus 0.125 for rumour cascades; n = 39/130/296 at MIN_STANCE = 15/8/5 under the lexicon, n = 30 for the classifier stream; S4–S6). A formal between-curve test replaces the side-by-side comparison: permuting rumour/non-rumour labels at cascade level (B = 500), the L2 distance gives p = 0.002 and the excursion difference 0.058 (p = 0.036, lexicon) and 0.104 (p = 0.002, classifier) (S8). Post-hoc power at the observed effect size is only ≈ 0.47–0.61, so these tests sit at the edge of what 30–39 control cascades deliver; jointly they argue against, without eliminating, the possibility that S-shaped curves arise mechanically. The control caveat stands: only 2.1% of non-rumour cascades pass the primary threshold, so the conclusion is "no dynamics in the analysable, stance-active non-rumour subset" (§3; §4.1).

Three implications follow. First, collective correction of falsehoods is real but intrinsically partial: even as the stream is exhausted, only a minority of expressed stances converges to denial — the ceiling quantifies the stubborn remainder (Lewandowsky et al., 2012; Walter and Tukachinsky, 2020; Ecker et al., 2022; Wood and Porter, 2019; Smith and Sørensen, 2000). This reading is restricted to false claims: the true-rumour curve also rises (K_m = 0.172), so its stream is evidently not truth-conducive and its ceiling is not evidence of failed correction. Second, unverified claims behave socially like false ones: their ceiling (0.314) sits nearer the false than the true value, consistent with unresolved claims attracting correction dynamics like debunked ones. Third, the ceiling is estimable but measurement-conditioned: on Weibo it ranges from 0.50 to 0.92 across denominator corrections, so absolute values should be compared only within a fixed pipeline; K_m can be tracked across topics, platforms, and time. Descriptive tracking of rumour-correction cascades exists

(Friggeri et al., 2014; Zubiaga et al., 2016, 2018); this design adds an estimable scalar with an inference apparatus.

Table 2. Fitted parameters and test statistics, by measurement variant (all fits multi-start).

Dataset / group	Knowledge axis / stance	n	K_m [95% CI]	R^2 (logistic)	ln(RSS ratio) point (boot. median)	logit win (B)	Placebo (B)
PHEME false	volume-V / human	39	0.372 [0.282, 0.478][a]	0.984	−0.105 (−0.066)	80% (500)	0.984 vs 0.954, k=8, p=0.009 (1000)
PHEME true	volume-V / human	80	0.172 [0.128, 0.227][a]	0.990	−0.062 (−0.022)	57% (300)	—
PHEME unverified	volume-V / human	65	0.314 [0.253, 0.378][a]	0.999	−0.371 (−0.084)	71% (300)	—
PHEME false	content-V / human	39	0.390 [0.305, 0.495]	0.981	−0.051 (−0.026)	67% (300)	0.981 vs 0.954, k=0, p=0.005 (200)
PHEME true	content-V / human	80	0.181 [0.133, 0.235]	0.979	−0.016 (+0.003)	49% (300)	0.979 vs 0.960, k=3, p=0.020 (200)
PHEME unverified	content-V / human	65	0.308 [0.256, 0.365]	0.994	+0.005 (+0.014)	47% (300)	—
CED rumour	volume-V / lexicon	329	0.619 [0.580, 0.949]	0.992	−0.087 (−0.093)	97% (300)	0.991 vs 0.910, k=0, p=0.005 (200)
CED rumour	volume-V / classifier	245	0.502 [0.467, 0.690]	0.994	−0.148 (−0.076)	81% (300)	0.994 vs 0.725, k=0, p=0.005 (200)
CED rumour	content-V / lexicon	329	0.795 [0.604, 1.000][d]	0.996	−0.003 (−0.001)	53% (300)	—
CED rumour	content-V / classifier	245	0.519 [0.475, 0.791]	0.997	−0.065 (−0.017)	71% (300)	—

Dataset / group	Knowledge axis / stance	n	K_m [95% CI]	R^2 (logistic)	ln(RSS ratio) point (boot. median)	logit win (B)	Placebo (B)
CED rumour, support-corrected (co-primary)	volume-V / lexicon, audit precision 0.18 / 0.10 / 0.05	329	0.823 [0.799, 0.843] / 0.875 [0.861, 0.894] / 0.919 [0.906, 0.936]	0.976 / 0.982 / 0.987	—	—	—
CED rumour, strict-support (co-primary)[f]	volume-V / assertive-support only	200	0.851 [0.827, 0.870]	0.980	—	—	—
CED non-rumour	volume-V / lexicon	39 (up to 296 at lower thresholds)	flat (fits pin at the K_m bound 0.02)[d]	excursion ≤ 0.07	(flat)	—	—
CED non-rumour	volume-V / classifier	30	flat	excursion ≤ 0.07	(flat)	—	—

Notes. [a] PHEME volume-V intervals are BCa (B = 2,000); other PHEME rows and all CED rows are thread-bootstrap percentiles (B = 300). Event-cluster bootstrap intervals for the PHEME volume-V rows: false [0.253, 0.499], true [0.131, 0.212], unverified [0.210, 0.362]; a paired event-cluster bootstrap (500 replicates, each sharing one resampled event set across both groups) gives P(false ≤ true) = 0.012; per-side resampling uses the 7/5/6 usable events. [b] $\ln(\text{RSS ratio}) = \ln\left(\frac{\text{RSS}_{\text{logit}}}{\text{RSS}_{\text{probit}}}\right)$, negative favours logistic; grid-invariant across 25/50/100-point grids (Supplementary Note S8); bootstrap median in parentheses. The unverified volume-V row illustrates why we report both: the point estimate (−0.371, a 31% RSS difference) is pulled by the curve's long flat tail, while the bootstrap median (−0.084) is far weaker — the 71% win rate, not the point estimate, is the representative summary. For the PHEME false row the two cell summaries are drawn from different archives: the bootstrap median is from the B = 300 archive and the win rate from the B = 500 archive (both archived; see Supplementary Note S8). [c] The leave-one-thread-out predictive comparison is demoted to

a diagnostic (§2.4): win rates 36% (PHEME false) and 51% (CED classifier); per-thread ΔSSE medians $+7.3\times10^{-4}$ and -2.7×10^{-5}. [d] CIs touching the K_m bound of 1.0 reflect genuine weak identification on upper-segment-only Weibo curves; bootstrap boundary mass per variant (K_m and V_0 pinning) is reported in §4.6 and Supplementary Note S8. Flat-control fits pin K_m at its lower bound 0.02, consistent with no estimable revision dynamics. [e] Placebo entries give observed R^2 vs permutation-null 95th percentile with (k+1)/(B+1) p values. Sensitivity of content-V rows to the de-duplication threshold (Jaccard 0.3–0.7, multi-start): CED K_m 0.791–0.840 (lexicon), 0.518–0.521 (classifier); PHEME ceilings vary ≤ 0.004 (Supplementary Notes S5, S8). Inclusion-threshold sensitivity and support-side correction details in Supplementary Notes S6–S8. [f] In the strict-support row the V_0 estimate sits at its lower bound (−0.60), reflecting weak identification of the onset under the strict-support restriction; the K_m interval is unaffected.

Table 3. Event × veracity distribution of PHEME threads (raw release counts / threads usable at ≥ 3 stance posts). The four events with ≥ 3 usable threads on both false and true sides (in italics) carry the exact event-level randomisation test.

Event	false (raw/usable)	true (raw/usable)	unverified (raw/usable)
charliehebdo	12/8	39/25	23/13
ebola-essien	2/2	0/0	0/0
ferguson	0/0	2/2	44/37
germanwings-crash	12/6	10/5	3/1
ottawashooting	10/8	35/19	13/7
prince-toronto	12/4	0/0	0/0
putinmissing	1/1	0/0	8/3
sydneysiege	13/10	51/29	7/4
Total	**62/39**	**137/80**	**98/65**

2.4 The sigmoid shape is established on Twitter; the logistic family is not identified

Per Proposition 3b, comparing link functions on belief curves identifies the underlying threshold distribution — with the Chambers–Cox caveat that such identification is asymptotically expensive and tail-driven (Chambers and Cox, 1967), and with the pooling

caveat of §2.1. Three lines of evidence do not agree; together they indicate that the family is not identified on these data. Throughout we use grid-invariant statistics — the log residual-sum-of-squares ratio $\ln\left(\frac{\mathrm{RSS}_{\mathrm{logit}}}{\mathrm{RSS}_{\mathrm{probit}}}\right)$ (negative favours logistic), the RSS percentage difference, and the bootstrap win rate — because information-criterion differences scale arbitrarily with grid size (S8).

In-sample fit leans logistic on the large sample, modestly. On the Weibo pooled curves the logistic fits better: ln(RSS ratio) = −0.148 (RSS difference 13.8%) with bootstrap win rate 81% (B = 300) under the classifier-filtered measurement, and −0.087 (8.3%) with win rate 97% under the lexicon measurement (Fig. 4a). On Twitter false rumours the logit wins 80% of replicates (B = 500; ln ratio −0.105, RSS difference 10.0%). Content-axis variants are smaller ($|\ln \text{ratio}| \leq 0.065$), and across measurement variants the win rate spans 47–97% — a lean, not a verdict.

Non-sigmoid saturating baselines tie the logistic where only the upper segment is observed. Adding four simpler baselines — linear, logarithmic, monomolecular, and Michaelis–Menten — under the same multi-start NLS + bootstrap protocol (B = 300) yields the sobering result of this section (Fig. 4c, S8). On both Weibo rumour curves the monomolecular model is indistinguishable from the logistic and slightly better (RSS ratios 0.889 and 0.864; logistic win rates 17% and 32%): on upper-segment-only data, an S-shaped and a purely concave saturating law are empirically interchangeable. On the Twitter false curve the two are essentially tied (RSS ratio 0.922; win rate 53%), while linear, logarithmic, and Michaelis–Menten baselines lose clearly (84–100%). The logistic wins outright only on the PHEME true and unverified curves. We therefore downgrade the Weibo shape claim: those curves establish monotone saturation with late-segment deceleration, consistent with the upper segment of a sigmoid.

The inflection of the Twitter curve is interior but the latent regime is usually not observed. The false-rumour inflection bootstraps to [0.078, 0.468] with 98.7% of mass inside the observed range; yet the pooled start falls below the logistic lower transition ($p_0 < 0.211K_m$) in only 34.7% of replicates, and, per §2.2, the apparent acceleration is largely the small-denominator artefact: after correction Twitter enters observation at $p_0/K_m \approx 0.83$, at or above the upper transition.

Out-of-sample prediction is at chance by construction. The per-thread ΔSSE (logit − probit) distribution is centred near zero on both platforms (Twitter false: $+7.3\times10^{-4}$, p = 0.108; Weibo classifier: -2.7×10^{-5}, p = 0.898), while between-thread heterogeneity dominates held-out error by an order of magnitude (RMS ≈ 0.32 against a curve excursion of 0.125). The design is powerless at realistic sample sizes; we demote it to a diagnostic (S8).

The asymmetry fit is suggestive, not decisive. A Richards curve nesting the logistic ($\nu = 1$) drives the asymmetry parameter to the boundary: $\hat{\nu} = 0.05$ in both cases, with profile intervals [0.05, 0.45] (Twitter false) and [0.05, 0.85] (Weibo classifier) that exclude $\nu = 1$. Three cautions demote this to a hint: the estimate is boundary-pinned; bootstrap replicates pin at the boundary 55.7% (Twitter) and 59.0% (Weibo); and on upper-segment-only curves ν and K_m are barely jointly identifiable. The fitted Richards curves imply third-derivative zeros at 0.081 and 0.691 of the ceiling, versus non-parametric change points 0.299/0.726 — compatible at the upper transition, inconsistent at the lower one. The hint points to Gompertz-type asymmetry.

Shape-specific placebo statistics. Because an R^2 gap can be opened by any monotone trend, we added two shape-specific statistics to the permutation test (S8): the Spearman correlation between windowed denial share and V, and the total excursion. Standardised gaps are z = 1.44 (Spearman) and z = 3.51 (excursion) on Twitter false rumours, and z = 1.66 /

9.39 on the Weibo classifier curve (Fig. 5b) — sequence-dependence is strongest in the excursion dimension.

Our summary: the data establish a saturating, monotone revision law with a ceiling, genuine sequence-dependence, and — on Twitter — an interior inflection; they do not identify the functional family, as Chambers and Cox predict at this sample size. Three conclusions are family-independent: a ceiling exists; the law is monotone and saturating; the curve encodes sequence, not just composition. We use the logistic as the working family; all family-dependent claims (including Proposition 4's thresholds) are conditional on that choice.

2.5 The structure is not an artefact of ordering, cumulation, labels, knowledge proxies, composition, or inclusion thresholds

Ten robustness checks guard the main claims (details in Supplementary Notes S2–S10; Fig. 5a–e). The confirmatory family is exactly two tests — the primary-variant ceiling contrast and the primary placebo — Holm-corrected (raw $p = 0.012$ and 0.009, adjusted 0.018; S11); all other tests are exploratory. Placebo. Permuting stance labels within threads destroys the curves: observed R^2 exceeds the null 95th percentile in every group ($p \leq 0.009$). The evidence is thus the observed–null gap, not raw R^2. Two stronger placebos ($B = 500$) decompose the effect: a cross-thread label shuffle leaves observed curves above the null 95th percentile in all four groups ($p = 0.002$–0.008), and block permutation preserves local order and remains below the observed curves ($p \leq 0.004$). De-cumulation. Windowed instantaneous shares (30%) preserve the rise (Twitter 0.31 → 0.46, $R^2 = 0.97$; Weibo classifier 0.43 → 0.53, $R^2 = 0.97$); the low cumulative start is largely a small-denominator product (0.103 → 0.529 after dropping $k = 1$–3 observations; Weibo 0.357 → 0.374). Stance measurement. A 360-item gold standard supports a two-stage pipeline: the lexicon proposes candidates (false-negative rate 1.25%) and a character n-gram logistic filter removes non-assertions (precision 0.734, recall 0.935, F1 0.822 under nested cross-validation). At that precision a quarter of

retained denials are plausibly non-denials, so K_m is biased upward — opposite to the support-side contamination. The support lexicon proved amplification-dominated (audit precision 0.10): correcting the denominator gives Weibo ceilings of 0.823/0.875/0.919, and a strict-support variant 0.851; the ceiling roughly doubles, so Weibo values are quoted as 0.50–0.92. Knowledge proxy. A content axis (character 3-gram Jaccard) leaves every qualitative conclusion intact and improves fit (Weibo classifier R^2 0.997; PHEME ceilings replicated within 0.02); a threshold sweep 0.3–0.7 moves the ceiling only between 0.791–0.840 (lexicon) and 0.518–0.521 (classifier; Fig. 5a). Time drift remains possible: de-duplication and the flat control constrain but do not exclude it. Composition. Three diagnostics address who and what enter p: user de-duplication shifts all ceilings by ≤ 0.02 (CED 0.619 → 0.604; PHEME false 0.372 → 0.359); the query-migration premise is rejected (query share rises rather than drains, and adding queries preserves the contrast, K_m = 0.295 vs 0.124); and the deny fraction drifts upward over V on Weibo (0.469 → 0.648), a drift that is itself the phenomenon but warns that the denominator is not a fixed population. Inclusion and pooling. Raising the Twitter minimum to 5 preserves the ordering (0.427 vs 0.258, ratio 1.7), so the headline ratio is quoted as 1.7–2.4; partial pooling leaves the contrast intact (shared-α fit α = 6.08, K_m = 0.360 vs 0.182; shrinkage erases it only at $\lambda \geq 10$, RSS +990%); and residuals under all three families are strongly autocorrelated (runs test $p < 10^{-4}$), reinforcing the family caveat.

2.6 A boundary condition: the response–uncertainty relation and the exchange rate

The framework predicts that the marginal response per unit knowledge follows $\frac{dp}{dV} = \left(\frac{\alpha}{K_m}\right) \cdot p(K_m - p)$ (Proposition 4iv). Pooled increments recover the predicted bell-shaped profile qualitatively: correlating pooled-curve increments with the $p(K_m - p)$ profile gives r = 0.410, p = 0.0034 (n = 49 increments, Weibo main curve). Two sharper tests, aimed at A1 rather than at curve shape, give a deliberately mixed picture.

Exponent law. Estimating the exponent in $\text{resp} \propto [p(K_m - p)]^\gamma$ by log–log regression on positive increments with $\bar{p} < K_m$ gives $\hat{\gamma} = 0.773$ on Weibo (iid SE 0.150; n = 47 increments); propagating within-cascade correlation and the generated-regressor nature of K_m by cascade-level block bootstrap (B = 300) moves the estimate to median 0.62 (SE 0.26, 95% interval [0.22, 1.26]), so $\gamma = 1$ sits at the upper edge and can be neither rejected nor confirmed. On Twitter the test is underpowered by design: $\hat{\gamma} = 0.45$ with n = 29 usable increments, post-hoc power 0.12 to distinguish $\gamma = 1$ from 0.45 (80% power would require n ≈ 304–400), whereas the pooled-increment design at the same n yields power 0.95 — the test is design-sensitive, and we report both (S8). A sub-one point estimate together with the boundary-pinned Richards asymmetry (§2.4) is suggestive but not decisive counter-evidence against the exact logistic law; plausible culprits include stance-label noise, burst-driven revision around external debunking events (Kleinberg, 2002), and heterogeneity of α across claims.

Segmented exchange rate — the ceiling-adjusted logit diagnostic. A direct diagnostic of A1 estimates α separately within segments of each pooled curve. The correct transform is the ceiling-adjusted logit $\ell(V) = \log\left[\frac{m(V)}{K_m - m(V)}\right]$, where m(V) is the pooled mean denial share: under the model $\ell(V) = \alpha(V - V_0)$ exactly (Proposition 5, S1), so segment slopes of ℓ on V are unbiased segment-specific estimates of α. The naive logit $\log\left[\frac{m(V)}{1-m(V)}\right]$ is not usable: its slope under a constant α is $\frac{\alpha(1-\sigma)}{(1-K_m\sigma)}$, which declines in V even when A1 holds exactly — an earlier draft made that transform error and read the mechanical decline as a violation of A1. Under the corrected transform the four-segment estimates are [2.41, 1.46, 2.00, 2.85] on Weibo (lexicon; fitted α = 2.02), [3.31, 2.38, 2.91, 4.21] (classifier; α = 2.86), and [4.87, 5.25, 3.63, 11.48] on Twitter false rumours (α = 5.14; the final segment rests on few grid points below K_m). No segment sequence declines monotonically (|Spearman ρ| ≤ 0.4), and each fluctuates around its fitted α: the segmented estimates are compatible with a constant

exchange rate, so A1 survives this sharp test — with the caveat that compatibility is weak confirmation, not proof (S1, S8).

3. Discussion

Summary. Across two platforms, two languages, and two sampling mechanisms, pooled collective belief revision behaves as a saturating, monotone function of accumulated knowledge: genuinely sequential (placebo $p \leq 0.02$; $p \leq 0.009$ for the primary volume-based analyses), not manufactured by cumulative averaging, stable under event-level clustering (leave-one-event-out ratio 1.98–2.41), and bounded by a ceiling that separates false from true information by a factor of ≈2 (1.7–2.4; paired event-cluster bootstrap $p = 0.012$). Three claims remain weaker: the logistic family (non-sigmoid baselines tie it on upper-segment-only curves; out-of-sample prediction is at chance; Richards asymmetry pins at its boundary), the ceiling's absolute value (0.50–0.92 across CED support-side corrections), and the latent regime, which is extrapolated. Given 39 false-rumour threads in 8 events (only 4 supporting an event-level exact contrast), the ceiling contrast and anatomy are an exploratory regularity, not a confirmatory estimate.

The ceiling as a measure of correctability — for false claims. $K_m \approx 0.37–0.39$ for Twitter false rumours, with no dynamics for non-rumours, quantifies a familiar question: how much correction can information buy? The continued-influence literature shows corrections rarely erase misinformation (Lewandowsky et al., 2012; Walter and Tukachinsky, 2020; Ecker et al., 2022); backfire effects prove elusive, suggesting structural stubbornness rather than reactance (Wood and Porter, 2019); social-learning theory shows rational populations can fail to learn even with unbounded evidence (Banerjee, 1992; Bikhchandani et al., 1992; Smith and Sørensen, 2000); opinion-dynamics models carry a near-isomorphic stubbornness parameter (Friedkin and Johnsen, 1990, 1999). Our result makes the shortfall a stable, estimable

parameter of the collective process, . Two qualifications matter. First, the reading applies to false claims only: the true-rumour curve also rises, so knowledge accumulation decouples from truth-tracking; the Charlie Hebdo event, where false and true ceilings coincide (0.162 vs 0.171), shows the ≈2-fold contrast is an average over heterogeneous evidence production, not a universal constant. Second, absolute values are measurement-conditioned: K_m is an audit metric only within a fixed, validated pipeline; with that caveat, it can be tracked across topics and periods to inform platform governance and fact-checking policy (Lazer et al., 2018; Vosoughi et al., 2018; Pennycook and Rand, 2021).

What the ceiling is not: three exchangeable mechanisms. The defective-threshold reading (Proposition 3d) is one of at least three mechanisms nearly exchangeable within an observation window: (i) a defective threshold distribution with large α; (ii) a regular distribution with small α; and (iii) Jensen attenuation — under A1's first-moment-only condition, aggregating stochastic evidence paths through the sigmoid produces an apparent ceiling below 1 even with no never-converters. Our simulation makes (iii) concrete: with $\tau = 1.5$, the aggregated path deviates by up to 0.061 and yields an apparent K_m of 0.939 against a true 1.0 (S8). Estimated K_m and ν therefore mix defect mass with Jensen attenuation; we report both readings rather than adjudicating. The defective reading remains valuable because it equals a dogmatic-prior subpopulation and is falsifiable in principle: in absolute evidence units, a plateau strictly below 1 excludes the small-α alternative, a test our normalised axis cannot run.

The upper-segment anatomy and the correction window. Proposition 4 partitions revision into latent, high-response, and consolidation segments with parameter-free boundaries (logistic-conditional; §2.4). Because both datasets enter observation late (Weibo $p_0/K_m \approx 0.71$–0.76; Twitter ≈ 0.83, §2.2), the data support an upper-segment anatomy with an extrapolated latent regime: we observe the top of the high-response window and a

consolidation plateau, never a latent plateau. The practical reading is logistic-conditional: debunking and authoritative disclosure should have maximal leverage while collective belief transits the central window (roughly 21%–79% of its attainable range), and sharply diminishing leverage outside it. If the response law instead follows the estimated sub-one exponent ($\text{resp} \propto [p(K_m - p)]^{\gamma}$, $\hat{\gamma} \approx 0.6$–0.8, §2.6), the response peaks below $\frac{K_m}{2}$ and the effective window sits lower. We offer the window as a family-conditional planning heuristic consistent with evidence that correction effectiveness varies with timing and design (Walter and Tukachinsky, 2020; Ecker et al., 2022), with the warning that after consolidation additional evidence is almost without effect. Corollary 4v remains a post-hoc consistency check; a genuinely falsifiable version awaits data with earlier observation.

Family identification: an honest retreat. Weibo pooled curves fit logistic better than probit in-sample (win rates 80–97%, grid-invariant), while non-sigmoid saturating baselines tie the logistic on those curves, held-out prediction is at chance by construction, and the best-fitting sigmoid is mildly asymmetric in a boundary-pinned sense. If thresholds are heavier-tailed than Gaussian — consonant with polarisation research (Del Vicario et al., 2016; Zollo et al., 2017) — the logistic remains the working model; but the family question is open, and Proposition 4's thresholds are logistic-conditional. What survives is the triple: a ceiling exists; revision is monotone and saturating; the curves encode sequence, not composition.

Relation to prior work. The logistic link is long established in psychometrics and econometrics (Rasch, 1960; McFadden, 1974); S-curves underpin diffusion modelling (Rogers, 1962; Bass, 1969); threshold aggregation is classical in sociology (Granovetter, 1978; Watts, 2002); MacCoun's BOP/BEAN models combine a logistic threshold form, an evidence axis, and an explicit ceiling close to ours (MacCoun, 2012, 2015); constant-drift evidence accumulation is standard in decision neuroscience (Gold and Shadlen, 2007; Glaze

et al., 2015); pooled information is the subject of collective-intelligence and prediction-market work (Hong and Page, 2004; Wolfers and Zitzewitz, 2004); incomplete social learning and stubborn agents are classical in economics and opinion dynamics (Banerjee, 1992; Bikhchandani et al., 1992; Smith and Sørensen, 2000; DeGroot, 1974; Friedkin and Johnsen, 1990, 1999); and the SEIZ model's Skeptic compartment is the epidemic-side precedent of a sub-one ceiling (Daley and Kendall, 1965; Jin et al., 2013). Two empirical antecedents — Zubiaga and colleagues' finding that false rumours attract denial while true rumours attract support (Zubiaga et al., 2016, 2018), and Friggeri and colleagues' tracing of Facebook rumour cascades and corrections (Friggeri et al., 2014) — are extended here from description to an estimable, cross-topic scalar K_m with an explicit knowledge axis. The contribution is therefore Proposition 4's differential properties repurposed as a family-discrimination device, plus a cross-platform, veracity-stratified ceiling-estimation design whose family-independent outputs survive both platforms and measurement generations.

Limitations. (i) Observational design: belief is shown to track accumulated information with saturating structure, not which items cause revision. Selective exposure and homophily, and external shocks, remain open non-causal accounts; placebos and de-cumulation constrain but cannot exclude them. (ii) Measurement validity: p is the share of expressed stances, not internal belief; posters are self-selected, the denominator's composition drifts, and user de-duplication changes K_m by ≤ 0.02 while the query-migration hypothesis fails. The Weibo support lexicon was amplification-dominated (audit precision 0.10), so the corrected pipeline is co-primary and absolute ceilings are measurement-conditioned (0.50–0.92). (iii) Family identification is unresolved: non-sigmoid baselines tie the sigmoid on upper-segment-only data, and residuals under all three families remain strongly autocorrelated (runs test $p < 10^{-4}$). (iv) Weak identification on upper segments: several Weibo fits pin parameters at bounds (V_0 at its lower bound in up to 37% of bootstrap replicates; 40.7% boundary mass on one K_m

variant). (v) Clustering and attrition: Twitter threads nest within 8 events (exact inference limited to G = 4, minimum p = 0.125); the Weibo non-rumour control rests on 39 analysable cascades (2.1% of the release), and the two CED groups enter through asymmetric mechanisms with labels of different validity than PHEME's. (vi) $\gamma = 1$ can be neither rejected nor confirmed once within-cascade correlation and generated-regressor uncertainty are propagated (§2.6); the Twitter test is underpowered (0.12). (vii) The lexicon ignores negation scope, the gold standard lacked a second coder, and the content axis is lexical, not semantic. (viii) The corpora date from 2009–2015; estimates should not be extrapolated to current platforms.

Outlook. Three extensions follow. First, equipping V with explicit semantic content (embedding-level de-duplication, knowledge-graph triples) connects the framework to that literature and enables cross-claim comparison of α — Proposition 4(iii) predicts that high-α domains flip fast while low-α domains host durable controversy. Second, the family question calls for extremely informative tails (Chambers and Cox, 1967) or experiments manipulating evidence strength, which would also test A1 sharply. Third, the early-entry corollary suggests an early-warning use, but requires the early segment neither dataset contains.

4. Methods

4.1 Data

PHEME / RumourEval-2017 (Twitter). We use the PHEME rumour dataset (Zubiaga et al., 2016) in its stance-annotated redistribution as RumourEval-2017 (SemEval-2017 Task 8) (Derczynski et al., 2017); “PHEME” is the primary name used in figures and tables. The release comprises 297 rumour threads across 8 breaking-news events (2014–2015), 4,519 tweets, thread-level veracity labels (137 true, 62 false, 98 unverified) and per-tweet stance annotations (support / deny / query / comment) produced under the protocol documented by

Zubiaga et al. (2016) and Derczynski et al. (2017). We analyse the training+development release only, excluding the task's test split: that split was released later under a separate protocol, its event composition differs, and including it would break the homogeneous annotation protocol and the clean event nesting on which our cluster-level inference depends. (A later SemEval-2019 Task 7 release (Gorrell et al., 2019) covers different data; a distinct PHEME veracity-classification release on figshare (doi:10.6084/m9.figshare.6392078) is not used here.) Timestamps were recovered from tweet objects and cross-checked against snowflake ID decoding.

CED (Sina Weibo). The CED dataset comprises 1,538 rumour and 1,849 non-rumour repost cascades (2009–2015) (Song et al., 2021; Chinese companion article: Liu et al., 2018). Rumour cascades were collected from Sina's rumour-reporting platform and carry platform-adjudicated debunking labels; non-rumour cascades were sampled by the original authors from ordinary public posts screened against the platform's rumour registry. Two validity asymmetries follow. First, selection into the rumour group requires user reporting and platform adjudication, so rumour cascades are naturally stance-active while non-rumour cascades carry no such requirement — the attrition mechanisms differ (§2.5). Second, CED labels are platform verdicts, not PHEME's human labels; the two label types are not validity-equivalent. Misclassification is conservative for our control conclusion: if some “non-rumour” cascades are unreported false claims, the control pool is contaminated toward rumour-like behaviour, and the observed flatness survives. Post counts (1,275,106 reposts, 768,551 with text and 506,555 bare forwards) are the authors' tally from the raw files, reproducible with the released code (Code availability; reconciliation and checksums in S8). Text-less reposts define cascade timing and volume but cannot carry stance, and never enter the stance stream or p's numerator or denominator. The two datasets and their analysis-ready samples are summarised in Table 1.

4.2 Belief operationalisation

For each thread/cascade, posts were ordered by timestamp. Stance-bearing posts are those labelled support or deny (Twitter: human annotation; Weibo: two-stage measurement, §4.3–4.4). The collective belief state at step k is the cumulative denial share $p_k = \frac{(\#\text{deny})_k}{(\#\text{deny}+\#\text{support})_k}$. Queries and comments enter neither numerator nor denominator (the query-migration diagnostic of §2.5 examines the consequences). A de-cumulated variant (30%-window moving-average denial share; 20%/40% widths in sensitivity analyses) is used in the robustness analysis of §2.5.

4.3 Chinese stance lexicon (stage 1)

Denial markers (refined version): 谣言, 造谣, 辟谣, 假消息, 假新闻, 虚假信息, 不实, 误传, 流言, 澄清, 举报, 别信, 不信, 勿信, 不要信, 别传, 骗子, 扯淡, 胡说, 编造, 捏造, 子虚乌有, 无稽之谈, 假图. Interrogative forms (真的假的, 真假, 是不是谣言, 真的吗) were excluded after a manual spot-check of the initial lexicon (precision ≈ 60%) revealed query–denial confusion via substring matching (假的 ⊂ 真的假的). Support markers: 是真的, 确实是, 确实, 属实, 证实, 千真万确, 求扩散, 扩散, 帮转, 转起, 都看看, 震惊, 气愤. The stage-1 lexicon flags 10,316 candidate denial posts (1.34% of the 768,551 text-bearing reposts). Version history of the lexicon and of superseded implementations is documented in Supplementary Note S4. We state plainly how the support list was built and why it failed: markers were assembled by the authors from high-frequency co-occurring terms in stance-active cascades and screened by reading a small convenience sample, without a formal construct-validity check; the inclusion of emotion words (震惊 “shocked”, 气愤 “outraged”) as support markers was a design-stage error — such words mark arousal, not endorsement — and it is the proximate cause of the 0.10 audit precision reported in §2.5. The support-side

correction pipeline (§2.5, §4.4) exists precisely because this flaw was caught by audit rather than by construction.

4.4 Supervised stance filter (stage 2)

Gold standard: 360 manually annotated reposts (280 lexicon-flagged, 80 unflagged; coding rules and data in S4). Annotations were produced by the authors; no second-coder subset was collected, so we report no inter-annotator reliability (a limitation, §3). The unflagged stratum places the lexicon false-negative rate at 1.25% (1 of 80), exact 95% CI [0.03%, 6.77%], so filtering applies to flagged candidates only. Features: character 1–4-gram TF–IDF (sublinear TF, min_df = 2); model: L2 logistic regression (C = 10). Thread-level rumour verification (Kochkina et al., 2018) is a different task; here each repost is classified independently. Because the production threshold (0.60) was previously chosen on the evaluation data, we re-evaluated under nested cross-validation (5 outer folds, threshold re-tuned inside each fold): precision 0.734 [0.715, 0.750], recall 0.935, F1 = 0.822, AUC = 0.824 [0.778, 0.871]; inner thresholds ranged 0.32–0.46, so the production threshold is conservative. Applied at 0.60, the filter retains 6,461 denial posts (62.6%). A planned Chinese sentence-encoder baseline was skipped because the offline environment could not install the required package, which we record rather than claim a comparison we did not run.

Support-side audit and correction. We sampled 100 support-flagged posts (fixed seed) and manually coded whether each explicitly asserts or endorses the claim's truth (strict rule: conditionals, interrogatives, and neutral amplification do not count). Audit precision: 0.10 overall (exact 95% CI [0.049, 0.176]; assertive-marker subset 0.11; amplification subset 0.08), against a raw support pool of 4,198 posts (deny pool 5,859). The implied denominator amplification factor is 1.60× [1.52×, 1.66×]. Two co-primary corrections (§2.5): (i) partial correction of the denominator to the audit-implied genuine-support count at audit precision 0.18/0.10/0.05 (K_m = 0.823/0.875/0.919); (ii) a strict-support variant (markers 是真的/确实

是/确实/属实/证实/千真万确; interrogatives and conditionals excluded) retaining 3,740 posts after overlap with the deny lexicon is removed (K_m = 0.851 [0.827, 0.870], n = 200, R^2 = 0.980). Audit data and coding rules are released with the reproduction package.

4.5 Knowledge operationalisation

Volume-based V (primary): normalised information progress $V_k = \frac{k}{n}$ within the thread's stance-post stream. Content-based V (robustness): posts are greedily clustered in temporal order by character 3-gram Jaccard similarity (threshold 0.5; multi-start sensitivity to thresholds 0.3–0.7 reported in Supplementary Note S8) after normalisation (URLs, @-mentions, boilerplate removed); V_k is the fraction of the thread's total unique information units introduced up to post k. The implementation stores 3-gram sets as frozensets of strings — no hash-derived identifiers — so clustering and all downstream estimates are deterministic and bit-identical across runs. Mean unique-content fraction: 74.8% (Weibo), 82.4% (Twitter).

4.6 Pooled estimation and model comparison

Each thread's (V, p) trajectory was linearly interpolated onto a common 50-point grid spanning [0.02, 1]; the pooled curve is the unweighted thread-wise mean (size-weighted pooling in S7). Grid points below a thread's first observed V = 1/n are left-clamped by np.interp to that thread's initial share $p_1 \in \{0, 1\}$, which mechanically depresses the pooled start — the small-denominator artefact quantified in §2.2. Three-parameter logistic $p(V) = K_m \cdot \sigma\big(\alpha(V - V_0)\big)$ and probit $p(V) = K_m \cdot \Phi\big(b(V - V_0)\big)$ models were fitted by non-linear least squares (Levenberg–Marquardt; $K_m \in [0.02, 1]$, $V_0 \in [-0.6, 1.5]$) with multi-start optimisation (36 starting points; the global RSS optimum is reported, single-start fits converging to inferior local optima on the Weibo curves, S7). Uncertainty: thread-level bootstrap (B = 300), refitting per replicate; PHEME primary rows additionally report BCa intervals (B = 2,000) because the percentile method under-covers for small, skewed

distributions; for Twitter, additionally an event-level cluster bootstrap (B = 500, at its resolution floor — few-cluster intervals under-cover (Cameron et al., 2008), hence the LOEO and exact-randomisation complements of §2.3), plus a paired variant in which each replicate resamples the 8 events once and both ceilings are estimated within the same event set. Boundary behaviour is reported rather than smoothed over: K_m pinned at its upper bound in 40.7% (content-V/lexicon), 2.3% (volume-V/lexicon), 4.0% (content-V/classifier) and 0% elsewhere of replicates; V_0 pinned at its lower bound in 37.0% (volume-V/lexicon), 23.0% (content-V/lexicon), 5.0% (volume-V/classifier), 5.7% (content-V/classifier) and 0% on all PHEME rows. Model discrimination uses grid-invariant statistics — the log RSS ratio $\ln\left(\frac{\mathrm{RSS_{logit}}}{\mathrm{RSS_{probit}}}\right)$, the RSS percentage difference, and the bootstrap win rate — because information-criterion differences scale with the arbitrary number of grid points, whereas ln(RSS ratio) is stable across 25/50/100-point grids (S8). Inclusion: threads with ≥ 15 stance posts (Weibo) or ≥ 3 (Twitter); threshold sensitivity in §2.5 and S6. Attrition: Twitter 297 → 296 threads with any stance post → 184 at ≥ 3; Weibo rumour 1,538 → 1,405 → 245 (classifier) / 329 (lexicon) at ≥ 15; non-rumour 1,849 → 1,146 → 30 / 39.

We are explicit about the status of each inferential statement. The paired-bootstrap $P(K_m_false \leq K_m_true) = 0.012$ is a one-sided tail probability of the bootstrap distribution — Monte Carlo resolution 1/501 with B = 500 over 8 event clusters — not an exact test. The exact sign-flip randomisation test (p = 0.25, G = 4) remains the only exact event-level statement, and we report the two side by side rather than letting the sharper number stand alone. S12 classifies every inferential statement by type.

4.7 Change-point localisation

Regime transitions were localised with a third-order Savitzky–Golay filter (window 11; Savitzky and Golay, 1964) and the extrema of the smoothed second derivative around the

half-ceiling crossing, giving 0.2988/0.7263 of the fitted K_m (bootstrap 95% intervals [0.079, 0.499]/[0.534, 1.005], B = 300), compared with the logistic (0.2113/0.7887) and probit (0.1587/0.8413) candidates and the Richards-implied transitions (0.081/0.691) as a directional self-consistency check; full procedure in S8.

4.8 Placebo test

Within each thread, the binary stance labels were permuted and the cumulative denial share recomputed from the permuted sequence (stance-post positions and marginal composition preserved); pooled curves were refit (B = 200; B = 1,000 for the Twitter volume-based variant). All replicates use the full census of eligible threads/cascades — no subsampling. p values follow the (k+1)/(B+1) convention with k explicit (k = 0 of B = 200 gives p = 0.005). Beyond R^2, two shape-specific statistics are computed on the same replicates: the Spearman correlation between windowed denial share and V, and the total excursion, each with its standardised gap (§2.4, S8). Superseded implementations (cumulative-p permutation, first-50-cascade subsampling) are documented in S6.

Two stronger placebo designs were added (B = 500 each; seed 42; statistics: pooled-curve logistic R^2, total excursion, and windowed-share Spearman correlation). (i) Cross-thread label shuffle: all binary stance labels within a group are pooled and reassigned uniformly across threads, preserving each thread's stance-post count and the group-level marginal while destroying sequence and per-thread composition. (ii) Block permutation: within each thread the label sequence is divided into contiguous blocks which are permuted, preserving local order while destroying long-range order; block sizes 3, $\lfloor n/5 \rfloor$, and $\lfloor n/3 \rfloor$ form the sensitivity set. All other conventions are unchanged.

4.9 Response–uncertainty relation and segmented exchange rate

For each thread, increments $\frac{\Delta p}{\Delta V}$ were paired with interval midpoints $\bar{p}$; the exponent γ in resp $\propto [\overline{p}(K_m - \overline{p})]^{\gamma}$ — the form given by Proposition 4(iv), replacing an earlier misspecified p(1 − p) form — was estimated by log–log regression on positive increments with $\bar{p} < K_m$. Because the increments share cascades and K_m is itself estimated, uncertainty is quantified by a cascade-level block bootstrap (B = 300) that resamples cascades, rebuilds the pooled curve, re-fits K_m by multi-start, and re-estimates γ; because roughly one in six replicates suffers a collapse of the design-point span ($\log[p(K_m - p)]$ range < 1.0, versus 1.60 observed) producing explosive $|\hat{\gamma}|$, we report the design-stable subset (median 0.62, SE 0.26, 95% CI [0.22, 1.26]) alongside the raw distribution. Post-hoc power to distinguish $\gamma = 1$ from $\gamma = 0.45$ was simulated for the thread-level design (power 0.12 at n = 29; 80% power at n ≈ 304–400) and the pooled-increment design (0.95 at n = 29), and both are reported (S8). The segmented-α diagnostic splits each pooled curve into 3 or 4 V-segments and regresses the ceiling-adjusted logit $\ell = \log\left[\frac{m(V)}{K_m - m(V)}\right]$ on V over grid points with $m(V) < K_m$; an earlier draft used the naive logit, whose mechanical decline under a constant α was misread as an early-segment violation of A1 — corrected here (S8, S11).

4.10 Partial pooling and residual diagnostics

Partial pooling (MAP analogue). With no probabilistic-programming backend offline, pooling was implemented as penalised maximum likelihood: a shared-α joint NLS fit with group-specific K_m and V_0 (PHEME: false/true/unverified; CED: lexicon/classifier), plus a shrinkage sweep adding $\lambda \Sigma_g (K_{m,g} - \bar{K}_m)^2$ to the joint RSS for $\lambda \in \{0, 1, 10, 100\}$. Residual diagnostics (residual series, runs test, lag-1 autocorrelation) appear in Supplementary Fig. S1 and S10.

4.11 Software

Python 3.12; numpy, pandas, scipy (curve_fit, stats), scikit-learn (TfidfVectorizer, LogisticRegression), matplotlib. Analysis pipelines 1–25 are available (see Code availability).

References

Banerjee AV (1992) A simple model of herd behaviour. Q J Econ 107:797–817

Bass FM (1969) A new product growth for model consumer durables. Manage Sci 15:215–227

Bikhchandani S, Hirshleifer D, Welch I (1992) A theory of fads, fashion, custom, and cultural change as informational cascades. J Polit Econ 100:992–1026

Cameron AC, Gelbach JB, Miller DL (2008) Bootstrap-based improvements for inference with clustered errors. Rev Econ Stat 90:414–427

Chambers EA, Cox DR (1967) Discrimination between alternative binary response models. Biometrika 54:573–578

Cox RT (1946) Probability, frequency and reasonable expectation. Am J Phys 14:1–13

Daley DJ, Kendall DG (1965) Stochastic rumours. IMA J Appl Math 1:42–55

de Finetti B (1974) Theory of probability. Wiley

DeGroot MH (1974) Reaching a consensus. J Am Stat Assoc 69:118–121

Del Vicario M et al (2016) The spreading of misinformation online. Proc Natl Acad Sci USA 113:554–559

Derczynski L et al (2017) SemEval-2017 Task 8: RumourEval — determining rumour veracity and support for rumours. In Proc SemEval-2017, pp 69–76

Ecker UKH et al (2022) The psychological drivers of misinformation belief and its resistance to correction. Nat Rev Psychol 1:13–29

Friedkin NE, Johnsen EC (1990) Social influence and opinions. J Math Sociol 15:193–206

Friedkin NE, Johnsen EC (1999) Social influence networks and opinion change. Adv Group Process 16:1–29

Friggeri A, Adamic L, Eckles D et al (2014) Rumor cascades. In Proc ICWSM. AAAI

Glaze CM, Kable JW, Gold JI (2015) Normative evidence accumulation in unpredictable environments. eLife 4:e08825

Gold JI, Shadlen MN (2007) The neural basis of decision making. Annu Rev Neurosci 30:535–574

Good IJ (1950) Probability and the weighing of evidence. Griffin

Gorrell G et al (2019) SemEval-2019 Task 7: RumourEval — determining rumour veracity and support for rumours. In Proc SemEval-2019, pp 845–854

Granovetter M (1978) Threshold models of collective behaviour. Am J Sociol 83:1420–1443

Hong L, Page SE (2004) Groups of diverse problem solvers can outperform groups of high-ability problem solvers. Proc Natl Acad Sci USA 101:16385–16389

Jaynes ET (2003) Probability theory: the logic of science. Cambridge University Press

Jin F, Dougherty E, Saraf P et al (2013) Epidemiological modeling of news and rumors on Twitter. In Proc SNA-KDD '13. ACM. https://doi.org/10.1145/2501025.2501027

Keynes JM (1921) A treatise on probability. Macmillan

Kleinberg J (2002) Bursty and hierarchical structure in streams. In Proc KDD, pp 91–101

Kochkina E, Liakata M, Zubiaga A (2018) All-in-one: Multi-task learning for rumour verification. In Proc COLING, pp 3402–3413

Lazer D et al (2018) The science of fake news. Science 359:1094–1096

Lewandowsky S et al (2012) Misinformation and its correction: Continued influence and successful debiasing. Psychol Sci Public Interest 13:106–131

Liu Z et al (2018) 社交媒体平台谣言的早期自动检测 [Early automatic detection of rumours on social media platforms]. 全球传媒学刊 5:65–80. (In Chinese.)

MacCoun RJ (2012) The burden of social proof: Shared thresholds and social influence. Psychol Rev 119:345–372

MacCoun RJ (2015) Balancing evidence and norms in cultural evolution. Organ Behav Hum Decis Process 129:68–82

McFadden D (1974) Conditional logit analysis of qualitative choice behaviour. In: Zarembka P (ed) Frontiers in econometrics, pp 105–142. Academic Press

Pennycook G, Rand D (2021) The psychology of fake news. Trends Cogn Sci 25:388–402

Rasch G (1960) Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research

Richards FJ (1959) A flexible growth function for empirical use. J Exp Bot 10:290–300

Rogers EM (1962) Diffusion of innovations. Free Press

Savitzky A, Golay MJE (1964) Smoothing and differentiation of data. Anal Chem 36:1627–1639

Smith L, Sørensen P (2000) Pathological outcomes of observational learning. Econometrica 68:371–398

Song C, Yang C, Chen H et al (2021) CED: Credible early detection of social media rumors. IEEE Trans Knowl Data Eng 33:3035–3047

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359:1146–1151

Walter N, Tukachinsky R (2020) A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? Commun Res 47:155–177

Watts DJ (2002) A simple model of global cascades on random networks. Proc Natl Acad Sci USA 99:5766–5771

Wolfers J, Zitzewitz E (2004) Prediction markets. J Econ Perspect 18:107–126

Wood T, Porter E (2019) The elusive backfire effect: Mass attitudes' steadfast factual adherence. Polit Behav 41:135–163

Zollo F et al (2017) Debunking in a world of tribes. PLoS ONE 12:e0181821

Zubiaga A et al (2016) Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE 11:e0150989

Zubiaga A et al (2018) Detection and resolution of rumours in social media: A survey. ACM Comput Surv 51:32

Figure legends

Figure 1. Framework. Data → information → knowledge V → belief $p = K_m\sigma\big(\alpha(V - V_0)\big)$; the two parameter-free transitions at $0.211K_m$ and $0.789K_m$ delimit the latent (extrapolated, unobserved), high-response, and consolidation regimes; the impact curve $\frac{dp}{dV}$ peaks at $\frac{K_m}{2}$.

Figure 2. Pooled collective belief-revision curves. (a) PHEME/Twitter by veracity (false n = 39, true n = 80, unverified n = 65); (b) CED/Weibo rumour cascades under the classifier-filtered stance measurement (n = 245) with the flat non-rumour control. Points: thread-mean on a common grid; curves: three-parameter logistic fits (multi-start). Light reference lines: solid = observed pooled start p_0; dashed = the logistic lower transition $0.211K_m$, marked "extrapolated" — the latent regime below it is model extrapolation, not observation, on both

platforms (§2.2). R^2 alone is not evidence of sigmoid structure: the same pipeline yields pooled $R^2 \approx 0.9$ under a pure-chance null, so the evidential content lies in the observed–null gap (§2.5), the flat control (§2.3), and de-cumulation robustness (§2.5). Residual diagnostics for panels (a) and (b) — strongly autocorrelated under every candidate family — are reported in Supplementary Fig. S1 (§2.5).

Figure 3. Revision-ceiling estimates by veracity (PHEME) and measurement variant (CED): point estimates with 95% thread-bootstrap intervals (BCa for the PHEME volume-V rows); PHEME rows additionally show event-cluster bootstrap intervals. Light grey bars under the PHEME false/true volume-V rows give the leave-one-event-out ranges (false 0.331–0.434; true 0.164–0.183); together with the inclusion-threshold sweep the false/true ratio spans 1.7–2.4 (§2.3, §2.5). Intervals quantify fit uncertainty only — as with R^2, they are not by themselves evidence of sigmoid shape (see Fig. 2 caption).

Figure 4. Family discrimination. (a) Bootstrap distributions of $\ln\left(\frac{\mathrm{RSS}_{\mathrm{logit}}}{\mathrm{RSS}_{\mathrm{probit}}}\right)$ (negative favours the logistic link) for the PHEME false-rumour curve (bootstrap win rate 0.80) and the CED rumour curve, classifier-filtered (0.81); the statistic is grid-invariant (§2.4). (b) Non-parametric change-point estimates 0.299/0.726 with 95% intervals against the logistic (0.211/0.789, dotted) and probit (0.159/0.841, dashed) candidate thresholds; Richards-fit-implied transitions are marked for reference (0.081, open diamond; 0.691, filled diamond) — compatible at the upper transition, inconsistent at the unobserved lower one. (c) Non-sigmoid baseline comparison: RSS ratio relative to the logistic fit (log scale; values below 1 fit better than logistic) for five alternative laws (linear, logarithmic, monomolecular, Michaelis–Menten, probit) on three curves (PHEME false; CED lexicon; CED classifier); logistic bootstrap win rates against the monomolecular baseline are 53%/17%/32% respectively — the tie is exact where only the upper segment is observed.

Figure 5. Measurement robustness. (a) Multi-start Jaccard-threshold sweep (0.3–0.7) of the CED content-axis ceiling: lexicon K_m declines 0.84 → 0.79, classifier stays 0.521 → 0.518 (the primary threshold 0.5 reproduces the Table 2 rows exactly). (b) Shape-specific placebo: null distributions of the windowed-curve excursion (within-thread stance permutation, B = 500) versus observed values — PHEME false p = 0.006; CED classifier-filtered p = 0.002, using the (k+1)/(B+1) convention with B = 500. (c) Window-width sensitivity (20%/30%/40%) of the de-cumulated excursion: 0.145/0.145/0.132 (PHEME false) and 0.112/0.105/0.092 (CED classifier). (d) Attrition funnels: PHEME threads and CED rumour/non-rumour cascades from release to primary inclusion (§4.6). (e) Support-side denominator correction: Weibo ceiling by assumed audit precision (1.00 → K_m = 0.619 uncorrected; 0.18 → 0.823; 0.10 → 0.875; 0.05 → 0.919), with the independent strict-support variant (0.851) marked — the two correction strategies are mutually consistent (§2.5).

Data availability

The PHEME/RumourEval-2017 dataset is publicly available (SemEval-2017 Task 8 release; Derczynski et al., 2017); the related PHEME veracity dataset is on figshare (https://doi.org/10.6084/m9.figshare.6392078). The CED dataset is publicly available (https://github.com/thunlp/Chinese_Rumor_Dataset; Song et al., 2021). Processed datasets generated in this study — including the 360-item stance gold standard, the 100-item support-side audit, all pooled curves, and all fitted parameters — are included in the versioned reproduction package accompanying this submission (analysis version 15) and will be deposited in a public DOI-minting repository (Zenodo) before publication.

Code availability

All analysis code (pipelines 1–25, with a README mapping scripts to figures and tables and documenting the version history of measurement refinements and superseded

implementations) is included in the versioned reproduction package accompanying this submission and will be deposited in a public DOI-minting repository (Zenodo) before publication.

Competing interests

The authors declare no competing interests.

Ethical Approval

This article does not contain any studies with human participants performed by any of the authors.

Informed Consent

Not applicable. This article does not contain any studies with human participants performed by any of the authors. Informed consent was therefore not required.

a THEORETICAL FRAMEWORK — belief as a saturating map of accumulated knowledge §2.1

DATA STREAM → **INFORMATION** → **KNOWLEDGE V** → **COLLECTIVE BELIEF** $p(V) = K_m \cdot \sigma(\alpha(V - V_0))$

timestamped posts accumulate | semantically de-duplicated evidence | normalised information progress, V ∈ [0, 1]

- **A1** · constant expected log-evidence α (the bridge)
- **P1** · odds-form Bayes ⇒ exact sigmoid
- **P2** · Gaussian bridge, equal covariance
- **P3** · heterogeneous agent thresholds ⇒ population sigmoid
- **P4** · three segments at $0.211 / 0.789\ K_m$

b DATA & MEASUREMENT §4

TWITTER — PHEME / RumourEval-2017

297 rumour threads · 4,519 human stance annotations
veracity: false / true / unverified
belief p = cumulative denial share within thread

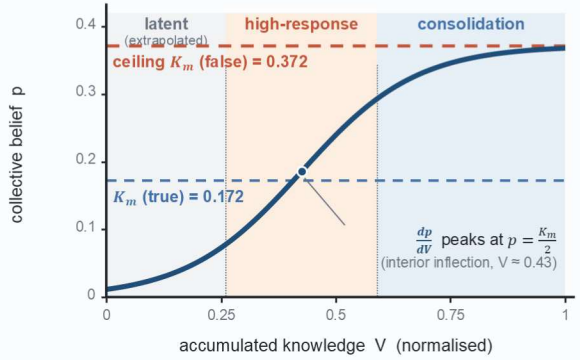
WEIBO — CED

3,387 repost cascades · 1.28 M reposts (2009–2015)
2-stage stance measurement: refined denial lexicon
→ supervised filter (360-post gold standard; audits)

KNOWLEDGE AXIS V

volume-based $V = \frac{k}{n}$ (primary)
content-based 3-gram Jaccard de-duplication (robustness)

c THE SIGMOID REVISION LAW §2.2–2.4



schematic — pooled-curve logistic fits, $R^2 \approx 0.98$ on both platforms

d KEY FINDINGS §2.2–2.6

- **$R^2 \approx 0.98$** — Saturating, monotone, sequence-dependent revision curves on both platforms
- **2.2×** — ceiling discriminates veracity: K_m false 0.372 vs true 0.172
  paired bootstrap P = 0.012; LOEO ratio ∈ [1.98, 2.41]
- **link** — Sigmoid shape established; logistic-vs-probit family not identified (Chambers–Cox)
- **γ ≈ 1** — Boundary law: marginal response ∝ $p(K_m - p)$; segmented exchange rate stable

10 robustness guards rule out ordering, cumulation, label, proxy, composition and inclusion artefacts (panel e)

e ROBUSTNESS — TEN GUARDS §2.5

placebo labels (P = 0.009)	**de-cumulation** windows	**deny-lexicon** audit	**support-side** audit	**content-based V**
numerator / denominator	**inclusion** thresholds	**per-thread** aggregation	**partial** pooling	**residual** diagnostics

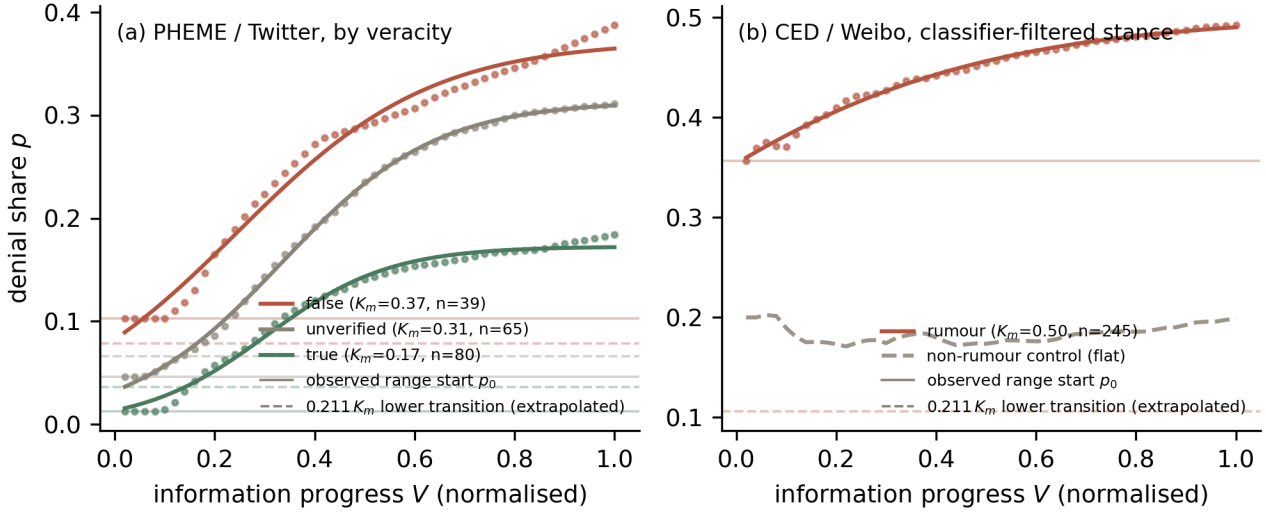
multiplicity discipline stated a priori · structure survives every guard

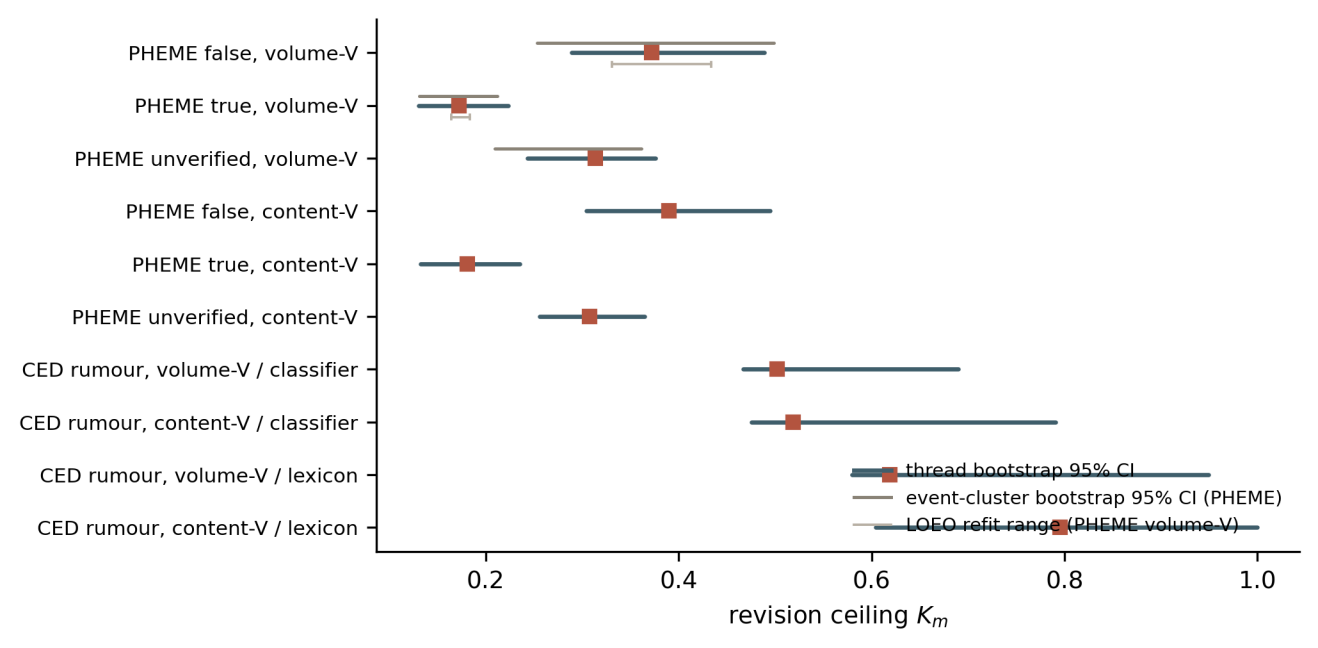
f IMPLICATIONS §3

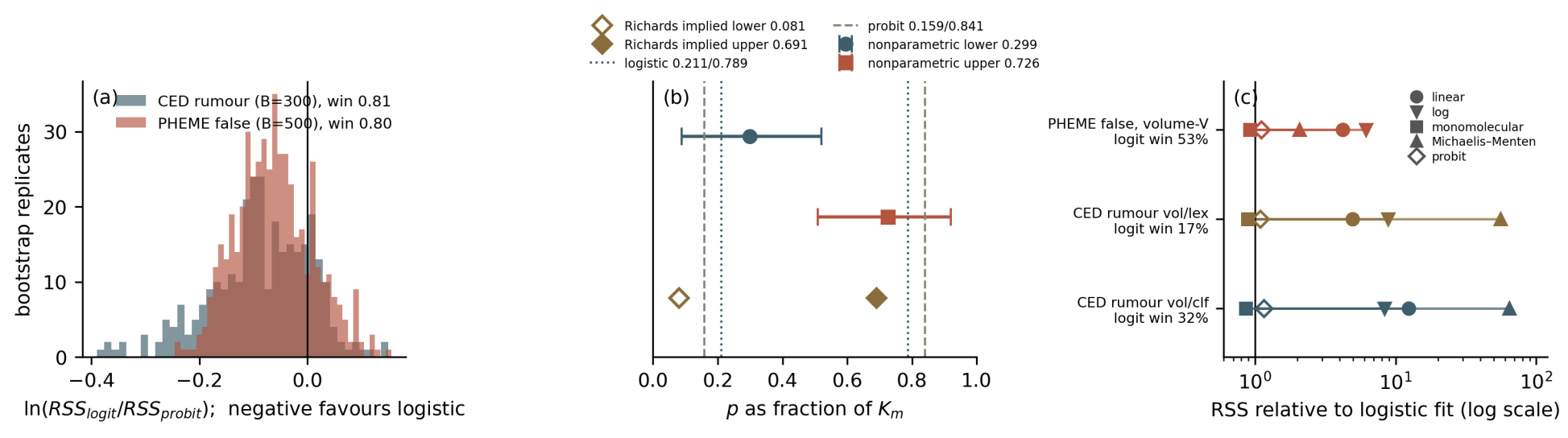
01 Correction is real but intrinsically partial — false-rumour denial saturates at $K_m \approx 0.37$, far below 1

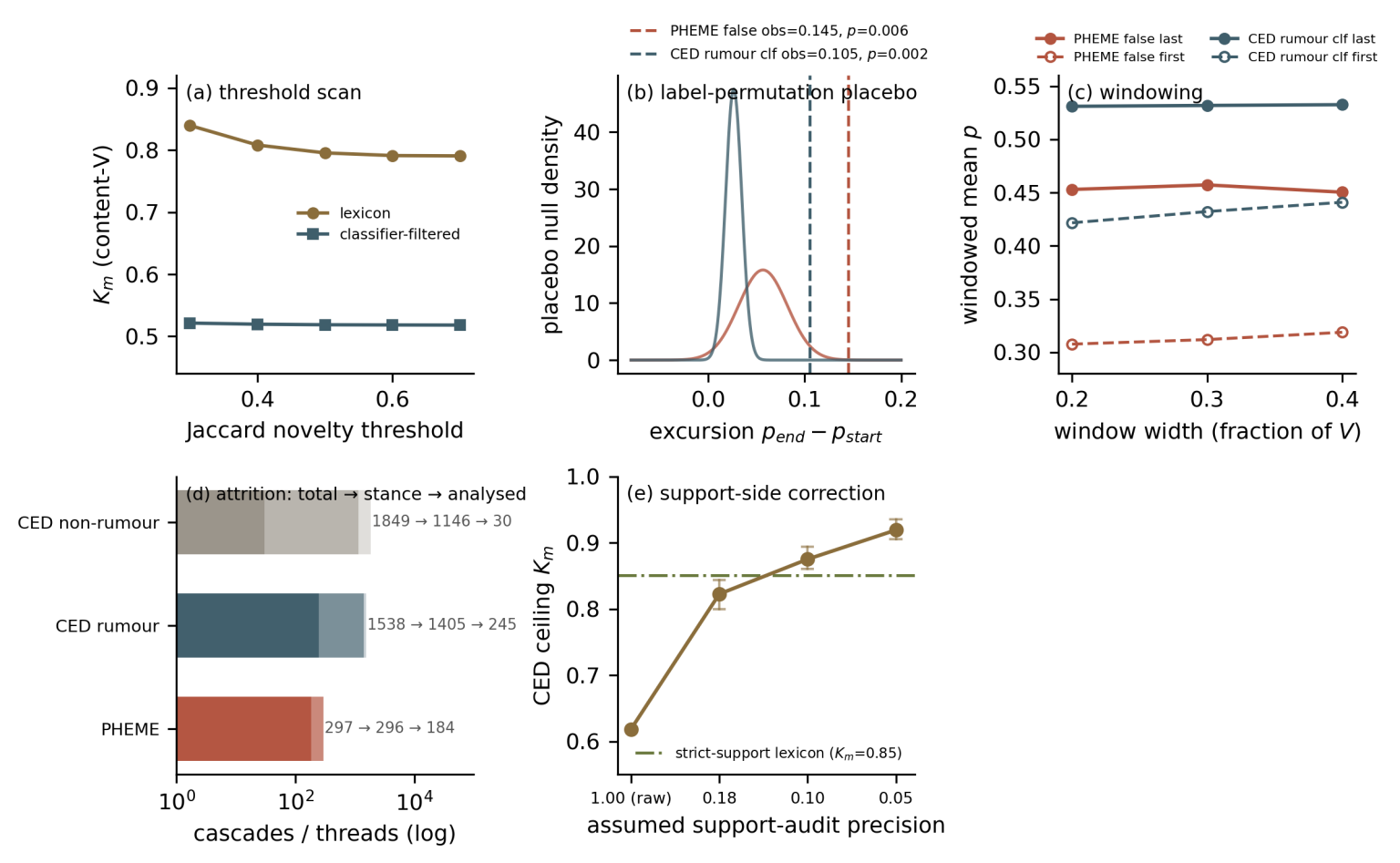
02 A measurable correction window — marginal response peaks at $p = \frac{K_m}{2}$: intervene early in the high-response segment

03 The ceiling K_m is a portable metric of collective correctability — comparable across platforms, languages and sampling mechanisms


(a) PHEME / Twitter, by veracity
denial share p
information progress V (normalised)
false (K_m=0.37, n=39)
unverified (K_m=0.31, n=65)
true (K_m=0.17, n=80)
observed range start p_0
0.211 K_m lower transition (extrapolated)
(b) CED / Weibo, classifier-filtered stance
information progress V (normalised)
rumour (K_m=0.50, n=245)
non-rumour control (flat)
observed range start p_0
0.211 K_m lower transition (extrapolated)


PHEME false, volume-V
PHEME true, volume-V
PHEME unverified, volume-V
PHEME false, content-V
PHEME true, content-V
PHEME unverified, content-V
CED rumour, volume-V / classifier
CED rumour, content-V / classifier
CED rumour, volume-V / lexicon
CED rumour, content-V / lexicon
thread bootstrap 95% CI
event-cluster bootstrap 95% CI (PHEME)
LOEO refit range (PHEME volume-V)
0.2
0.4
0.6
0.8
1.0
revision ceiling K_m


(a)
CED rumour (B=300), win 0.81
PHEME false (B=500), win 0.80
bootstrap replicates
0
10
20
30
−0.4
−0.2
0.0
ln(RSS_logit/RSS_probit); negative favours logistic
Richards implied lower 0.081
Richards implied upper 0.691
logistic 0.211/0.789
probit 0.159/0.841
nonparametric lower 0.299
nonparametric upper 0.726
(b)
0.0
0.2
0.4
0.6
0.8
1.0
p as fraction of K_m
(c)
linear
log
monomolecular
Michaelis–Menten
probit
PHEME false, volume-V
logit win 53%
CED rumour vol/lex
logit win 17%
CED rumour vol/clf
logit win 32%
10^0
10^1
10^2
RSS relative to logistic fit (log scale)


(a) threshold scan
K_m (content-V)
0.9
0.8
0.7
0.6
0.5
lexicon
classifier-filtered
0.4
0.6
Jaccard novelty threshold
PHEME false obs=0.145, p=0.006
CED rumour clf obs=0.105, p=0.002
(b) label-permutation placebo
placebo null density
40
30
20
10
0
0.0
0.1
0.2
excursion $p_{end} - p_{start}$
PHEME false last
PHEME false first
CED rumour clf last
CED rumour clf first
(c) windowing
windowed mean p
0.55
0.50
0.45
0.40
0.35
0.30
0.2
0.3
0.4
window width (fraction of V)
(d) attrition: total → stance → analysed
CED non-rumour
1849 → 1146 → 30
CED rumour
1538 → 1405 → 245
PHEME
297 → 296 → 184
10^0
10^2
10^4
cascades / threads (log)
(e) support-side correction
CED ceiling K_m
1.0
0.9
0.8
0.7
0.6
strict-support lexicon (K_m=0.85)
1.00 (raw)
0.18
0.10
0.05
assumed support-audit precision