# Decentralized Source Moderation at Scale: A Computational Analysis of Source Reliability Evaluation, Crowd Competence, and the Governance of Wikipedia's Deprecated Sources List

**Keywords:** Wikipedia, Decentralized Governance, Disinformation, Wisdom of Crowds, Content Moderation.

## 1 Background and Motivation

Wikipedia is among the most widely consulted reference resources in the world, with billions of monthly readers relying on it across nearly every domain. Its reliability depends on a mixture of factors, such as the accuracy of individual articles, but more interestingly also on the rigour of the source evaluation infrastructure underpinning them. Wikipedia delegates this evaluation entirely to volunteer editors, so that contrary to centralized platforms such as Facebook, X and Youtube, there is no corporate trust and safety team, no algorithmic classifier, and no paid fact-checkers. The Deprecated Sources List and its related reliable sources noticeboards (RSN) are key institutional mechanisms through which the community governs information quality. Together, they form an open deliberative forum where editors collectively determine which sources are reliable and which should be deprecated or blacklisted. Once deprecated, a source's citations become structurally detectable and removable across the encyclopaedia.

The question whether this decentralised moderation can achieve expert-competitive reliability assessments at scale has recently become more important to consider. The Digital Services Act designates the Wikimedia Foundation as a very large online platform, making robust source evaluation a legal as well as an editorial concern within their task to mitigate systemic risks on their platform. From the political angle, there has also been new interest in Wikipedia, as The U.S. House Committee on Oversight has launched an investigation into alleged coordinated manipulation of Wikipedia articles by foreign actors, creating external demands for centralized accountability that sit in direct tension with Wikipedia's community-autonomy model. Meanwhile, X's Community Notes and Meta's 2025 decision to replace contracted fact-checkers with community notes reflect a broader industry shift towards decentralized fact-checking systems that makes Wikipedia's decades-long track record a uniquely valuable empirical reference point.

## 2 Theoretical Framework

The analysis is situated within two competing frameworks. The wisdom-of-crowds literature [1,2] indicates that aggregated independent judgements outperform individual experts when the context is decentralized, independent, contains diversity of opinion, and can be aggregated. The RSN approximates these conditions structurally with its decentralized and deliberative structure. Against this, institutional logics theory [3] holds that organizational fields generate normative pressures that may undermine independence. If editors cluster ideologically or vote in line with political priors rather than source quality, the crowd's wisdom is compromised. These frameworks produce competing predictions that are empirically tested against the full RSN archive.

## 3 Data and Methods

This paper introduces a novel dataset of the complete RSN archive: 47,417 editor comments across 2,361 discussions linked to 473 unique sources spanning 2006–2025. This constitutes one of the first large-scale longitudinal bodies of Wikipedia's source reliability deliberations. A core methodological contribution is the application of a large language model (Claude Sonnet) to classify the political lean of each source, a task conventionally requiring costly manual coding. This outcome has been validated against AllSides and MBFC ratings. External credibility measures (factual accuracy scores, MBFC ratings, a Lewandowsky misinformation index) were merged with Wikipedia's reliability labels and reduced to a composite quality index.[4] For the discussions containing formal votes, a rule-based extractor identified vote-coded comments aggregated into per-discussion vote counts and plurality verdicts. Editor-level analyses drew on a bipartite editor–source co-discussion network (157 editors with ≥3 discussions), producing lean-bias scores, centrality measures, and community memberships via the Louvain algorithm.

## 4 Key Findings

**External quality aligns with community judgement**. Wikipedia's reliability labels correlate strongly with all three external quality metrics (Spearman r = 0.56–0.65), and the composite quality index explains 68.4% of variance in the ordinal reliability scale — supporting the crowd-wisdom interpretation. State-funded media from authoritarian contexts score significantly lower on both external metrics and Wikipedia labels than Western commercial or public-service broadcasters.

**Three structural blind spots qualify this finding**. First, domain boundary effects: state media receives a Wikipedia rating gap disproportionately wider than its expert quality deficit alone predicts, while platforms and UGC sources are rated almost as harshly as state media despite meaningfully higher expert scores. This suggests that Wikipedia applies content-model criteria that external raters do not penalise equally. Second, concentration of influence: just ten editors appear in 150–307 discussions each, and 17.4% of editors account for the majority of deliberative activity, structurally compromising the independence condition that wisdom-of-crowds theory requires. This may be problematic, as high network centrality of this kind can cause crowd wisdom to fail by amplifying existing bias rather than correcting it.[5] Third, a composition-driven lean gap: after controlling for expert quality, left-leaning sources receive Wikipedia reliability ratings that are more favourable than their quality scores predict, while centre and right-leaning sources are treated comparably to one another. On the continuous AllSides scale, each one-step rightward shift predicts a 0.21 wiki_ord reduction ($p < 0.0001$), yielding a cumulative left-to-right disadvantage of 0.83 wiki_ord units, approaching the distance between two full reliability categories. This gap is robust across AFM/MBFC and AllSides classifications and survives propensity matching, bootstrap resampling, and ordered logit specifications. This is consistent with the known left-of-centre demographic skew of the Wikipedia editing community.[6]

**Deliberation is asymmetric and geopolitically concentrated**. 63% of comments are neutral or procedural, 22% are explicitly pro-reliability, and only 10% are anti-reliability. This indicates that discussions are predominantly deliberative rather than adversarial. Deprecated source discussions are the most hostile in tone (mean sentiment = 0.033 vs. overall mean = 0.068). The Israel-Palestine conflict dominates geopolitical content, appearing in 3.1% of all comments — more than double the Russia-Ukraine share (1.6%) — and discussion spikes align closely with real-world event timelines, suggesting the RSN responds to broader epistemic crises rather than operating as a self-contained bureaucratic routine.

## 5 Contribution and Implications

This paper provides one of the first large-scale computational audits of Wikipedia's source evaluation process, simultaneously documenting its expert-competitive strengths and structurally predictable limitations. Theoretically, it shows that crowd wisdom and institutional contestation coexist at different levels of analysis. Methodologically, it demonstrates how LLM-based lean classification and rule-based vote extraction can scale credibility research beyond what manual coding permits. In the next steps, this research will be extended to individual vote-level data to enable editor-level behavioural inference beyond the observed community-level patterns. The findings have implications for platform regulators assessing decentralized moderation as an alternative to corporate content enforcement, and for researchers using Wikipedia reliability labels in downstream studies.

## 6 References


1. Davis-Stober, C. P., Budescu, D. V., Dana, J. & Broomell, S. B. When is a crowd wise? *Decision* 1, 79 (2014).
2. Surowiecki, J. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. (Doubleday Books, 2004).
3. Thornton, P. H., Ocasio, W. & Lounsbury, M. *The Institutional Logics Perspective: A New Approach to Culture, Structure and Process*. (Oxford Academic, Oxford, 2012).
4. Lin, H. *et al.* High level of correspondence across different news domain quality rating sets. *PNAS nexus* 2, pgad286 (2023).
5. Becker, J., Porter, E. & Centola, D. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences* 116, 10717–10722 (2019).
6. Greenstein, S. & Zhu, F. Is wikipedia biased? *American Economic Review* 102, 343–348 (2012).