

Title: Eligibility, search, and coding protocol appendix for “Professional and community-based fact-checking show different strengths, but neither performs strongly across trust, scalability, and impact”

Authors: Petter Bae Brandtzaeg (1), Silje Susanne Alvestad (2,3), Jonas R. Kunst (4,5), Asbjørn Følstad (6)

Date: July 30th, 2026

Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix C: Eligibility, search, and coding protocol

Review design

This study used a PRISMA-guided systematic review design to identify and synthesize empirical studies of human-driven professional and community-based fact-checking on social media.

Databases and search period

Searches were conducted in ACM Digital Library, Web of Science, Scopus, and JSTOR. Eligible publications were published from 2022 through 2025.

Search strategy

The search strategy comprised three conceptual groups: misinformation, social media platforms, and fact-checking or verification. Terms within each group were joined with OR, while the three groups were joined with AND.

(“misinformation” OR “disinformation”) AND (“social media” OR “platform*” OR “Twitter” OR “X” OR “Facebook” OR “WhatsApp”) AND (“fact-checking” OR “fact checking” OR “fact-check*” OR “misinformation correction” OR “Community Notes” OR “Birdwatch” OR “community fact-checking” OR “professional fact-checking” OR “crowdsourced verification” OR “peer fact-checking” OR “expert verification”)

Trust, scalability, and impact were analytical categories and were not required search terms.

Inclusion criteria

1. Peer-reviewed journal article or accepted full conference proceeding.
2. Empirical study.
3. Examined professional and/or community-based human fact-checking.
4. Fact-checking occurred on or was directly connected to social media.
5. Reported evidence relevant to trust, scalability, or impact.
6. Published from 2022 through 2025.

Exclusion criteria

7. Not peer-reviewed.
8. Not empirical.
9. Did not examine professional or community-based human fact-checking.
10. Did not concern social media.

11. Focused only on automated detection, media literacy, or general content moderation without examining human or institutional fact-checking.
12. Examined crowd accuracy judgments rather than deployed platform fact-checking with public annotations or visible contributor participation.

Screening

Records were screened by title and abstract and then by full text against the eligibility criteria. The search yielded 94 unique records. Twenty-five full texts were assessed for eligibility, four were excluded at the full-text stage, and 21 studies were included in the review.

The aggregate screening totals reported here should be read together with the original screening log, which provides the record-level inclusion and exclusion decisions.

Coding and synthesis

Three researchers read the included studies independently. Their interpretations were discussed collectively and refined through comparison until agreement was reached. Formal intercoder reliability was not calculated because the evidence was methodologically heterogeneous and the ratings required context-sensitive judgment.

Each study was assessed for trust, scalability, and impact. The ratings were interpretive summaries of the strength, clarity, and consistency of the evidence reported within each study; they were not standardized quantitative effect measures or assessments of overall study quality. Table C1 defines the thresholds used in the synthesis.

Table C1. *Coding framework for trust, scalability, and impact.*

Rating	General threshold	Evidence characteristics
High	Strong, consistent evidence across contexts; clear patterns of effectiveness, reach, and/or credibility	Clear author claims, multiple sources, and/or converging data (e.g., behavioral, perceptual, platform analytics). Effects generalize well across groups and/or settings.
Moderate	Conditional or mixed evidence; effectiveness depends on context, user group, and/or platform	Qualifiers present (“some”, “often,” “under certain conditions”); partial and/or domain-specific effects; trust/scalability/impact varies by audience and/or content.
Low	Weak, inconsistent, and/or negative evidence; minimal reach, trust, and/or behavioral effect	Lack of measurable outcomes; evidence of failure, disengagement, low trust, and/or structural bottlenecks; authors report limitations and/or ineffectiveness.
N/A	Insufficient evidence to assess the dimension	The study did not report evidence relevant to the dimension. N/A does not indicate weak or low performance.