Policy Journal of Social Science Review

Online ISSN 3006-4635

Print ISSN 3006-4627

Vol. 3 No. 11 (2025)

AI-Driven Fact-Checking and Disinformation: Opportunities, Limitations, and Implementation in Newsrooms

[1]Dr. Aniqa Ali

[1]*Assistant Professor, Department of Media and Communication Studies, International Islamic University, Islamabad*

[1]aniqa.naseer@iiu.edu.pk

Abstract

AI-generated texts and very fast platforms put a strain on newsroom verification. We introduce a guardrailed provenance-first fact-checking system integrating sparse-dense hybrid retrieval, 360-ranks cross-encoder, calibrated veracity, and human-in-the-loop review. It uses source whitelisting, time filters, adversarial defenses, and citation integrity (URL, timestamp, snippet) and presents uncertainty and rationale summaries with a CMS plugin. The system with a temporally held-out newsroom test set (N=1,500 claims) in evaluation gives Micro-F1 0.82, Macro-F1 0.79, AUROC 0.89, NDCG 10 0.82, p95 latency 3.6 s, hallucination rate 3.7 and citation correctness 95.8.

AI assistance in a within-subjects newsroom pilot (N=32) led to a decrease in task time of about 29%, a decrease in workload and an increase in the proper use of it, with desk-level throughput improvement (+82% claims/hour; +60% stories/day) and a reduction in the number of escalations. Error analysis indicates proximity errors and lack of evidence in breaking news. We talk about governance, auditability and safe learning through the feedback of the editors. Findings reveal that AI complements, but does not substitute, editorial judgment with accuracy.

Keywords: Fact-Checking; Disinformation; Retrieval-Augmented Generation (RAG); Provenance; Hybrid Retrieval; Human-In-The-Loop; Calibration And Trust.

Article Details:

Received on 15 Oct 2025

Accepted on 03 Nov 2025

Published on 05 Nov 2025

Corresponding Authors*:

Policy Journal of Social Science Review
Online ISSN 3006-4635
Print ISSN 3006-4627
Vol. 3 No. 11 (2025)

Introduction

The speed and amount of digital content have changed the course of propagation (mis)information. The time interval between production and massive exposure (days) is condensed into a few minutes by social feeds, messaging applications, and short-video platforms. In the case of newsrooms that work with the pressure of the deadline, this compression increases the tension: one unsubstantiated post can spread exponentially across the platforms before human fact-checkers can even take a break to see it. Simultaneously, automated verification has reached a high level: modern NLP and vision-language systems are able to retrieve evidence, evaluate support/refutation, and even write rationales behind a ruling (Guo et al., 2022; Aly et al., 2021). FEVEROUS and end-to-end datasets of multimodal fact-checking have pushed the task out of sentence-based checks, into realistic and messy cross-media conditions that are closer to newsroom work (Aly et al., 2021; Yao et al., 2023).

Generative models bring in fresh promises and threats. Large language models (LLMs) have the ability to summarize long documents, make inconsistent statements consistent, and generate attribution-rich copy in a few seconds, and also, they can hallucinate missing facts or misassign authorship and omit important disclaimers (Ji et al., 2023). Assessment suites including TruthfulQA demonstrate that when powerful models are deployed, it is not always clear that they are generating falsehoods that are popular, thus pointing to the necessity of verifiable grounding in journalistic settings (Lin et al., 2022). The two approaches, namely retrieval-augmented and research-and-revise, seek to address this gap by compelling models to access evidence and update outputs to be attributed (Asai et al., 2023; Gao et al., 2023). Simultaneously, the governance-oriented instruments, e.g., watermarking, aim to assist the provenance processes in distinguishing between the human- and AI-generated text, which can be beneficial in the context of rumor-trace tracking or the implementation of the AI disclosure regulations (Kirchenbauer et al., 2023).

In spite of the developments, the deployment of newsrooms is still not that easy. Verification involves more than accuracy of classification; it involves trained retrieval, clear justifications, variety of sources, checking on timeliness, and a design that fits the day to day routine of the editors and reporters. Furthermore, verification is becoming multimodal: textual assertions are usually accompanied by screen shots, graphs and short video clips. Multimodal pipelines are shown to be chainable in evidence retrieval, stance detection, and explanation generation, although in practice, they are significantly underperforming on ambiguous, low-resource, and quickly-changing stories (Yao et al., 2023). In the meantime, the scope of newsroom-safe evaluation must not solely acknowledge the excellence of the statement but to also mention attribution the quality of all factual statements being able to be provenanced (Rashkin et al., 2023).

We report: (i) dataset curation (reflecting newsroom-realistic claims) (ii) a model pipeline (as a generative model of retrieval-augmented generation) with research-and-revise editing (and explicit attribution checks); (iii) a two-month newsroom pilot (integrating our tooling into pitch meetings and copy desks); and (iv) an evaluation model (that jointly measures evidence recall/precision, claim-level accuracy, decision latency, attribution completeness, perceived usability). Based on the latest development in retrieval-augmented generation and post-hoc revision of the attribution, our pipeline is constructed with the goal of being able to support multimodal inputs after end-to-end fact-checking datasets (Yao et al., 2023).

Policy Journal of Social Science Review
Online ISSN 3006-4635
Print ISSN 3006-4627
Vol. 3 No. 11 (2025)

Newsrooms are experiencing an increase in the speed-accuracy gap: information disseminated faster than can be verified by humans, and the existing AI systems are not yet capable of being factual, attributive, and able to integrate well with workflow Delays (Guo et al., 2022; Lin et al., 2022; Rashkin et al., 2023). We seek to quantify the speed and accuracy of fact-checking that can be achieved with AI without losing attribution and editorial discretion; describe trade-offs between accuracy, latency and usability; and suggest governance controls (attribution checks, audit logs, watermark-aware provenance) that can be appropriate in newsrooms (Gao et al., 2023; Kirchenbauer et al., 2023). This work aims at text-based verification including image evidence (links, screenshots, social posts) and is interested in English-language and daily-news timelines sources of open-web and does not address deep fake video forensics or closed-source platform-data. The paper provides practical evidence to editors and toolmakers on how to use AI to responsibly fact-check at scale through the combination of dataset curation, a verifiability-first pipeline, a newsroom pilot, and a multidimensional evaluation framework (Aly et al., 2021; Asai et al., 2023; Yao et al., 2023; Ji et al., 2023).

Section 2 (Related Work) is a literature review of automated fact-checking (data sets, retrieval, verification, explanation), LLM hallucination and factuality, attribution-focused evaluation, and governance tools. Section 3 (Methodology) explains our corpus, annotation procedures, and end-to-end pipeline (retrieval, claim verification, attribution audit and human-in-the-loop UI). Section 4 (Results) is a reporting of intrinsic (accuracy, evidence quality, attribution completeness) and extrinsic newsroom (latency, editor workload, usability) metrics. In section 5 (Discussion), failure cases (temporal drift, sparse evidence, conflicting sources), ethical guardrails (transparency, bias, disclosure) and operationalization are considered. Section 6 (Conclusion) draws conclusions and work to be done.

LITERATURE REVIEW

The process of automated fact-checking (AFC) is usually broken down into four mutually reinforcing subtasks, namely: (i) claim identification/extraction, discovering statements that prove liable to check; (ii) evidence retrieval, locating potentially corroborating sources; (iii) veracity prediction, making predictions about whether a given evidence fragment suffers or refutes a claim or is irrelevant to it (Guo et al., 2022; Hardalov et al., 2022). Recent surveys highlight the fact that improvement has been made through the collaborative modeling of these steps using transformer topologies, whereas practical newscast pipelines continue to be characterized by modular structures through which human intervention between steps can be made (Guo et al., 2022).

The field has moved past initially English-only datasets and metrics to those that represent evidence provenance and pro forma reasoning. FEVEROUS (FEVER successor) incorporates table-text reasoning and asks systems to provide supporting snippets, which allows it to do metrics like FEVER/FEVEROUS score which is a combination of label accuracy and evidence adequacy (Aly et al., 2021). Such knowledge-intensive standards as KILT combine set of tasks (fact-checking, ODQA, entity linking) over a shared corpus, and in which provenance-conscious scoring is necessary to entice retrievers and generators to provide justifications based on citable evidence (Petroni et al., 2021). It is also important to domain and language diversify: X-FACT is aimed at cross-lingual claim verification, and it should be emphasized that the systems trained on English headlines fail to transfer to other languages (Gupta and Srikumar, 2021). Together these datasets standardize precision/recall/F1 by part and introduce task-specific metrics which make

Policy Journal of Social Science Review
Online ISSN 3006-4635
Print ISSN 3006-4627
Vol. 3 No. 11 (2025)

incorrect labels rewarded without sufficient citations, which is essential in newsroom use. In order to address this issue, it is necessary to differentiate between the two categories.<|human|>To combat this problem, the two categories are to be distinguished.

The idea of stance detection has grown out of textual entailment into a multi-evidence multi-document system. Results suggest that source-claim discourse, uncertainty, and pragmatic cues (e.g., hedging) are better modeled, although there is a weak generalization across outlets and topics (Hardalov et al., 2022). Whereas claim-evidence pairs are targeted by AFC, platform level signals, including propagation, networks and actors are the focus of disinformation detection. Heterogeneous graphs, which represent users, posts, and publishers, are also notable in identifying deceptive stories as they encode patterns of diffusion and community infrastructures that are related to low-credibility campaigns (Phan et al., 2023). The pros, including the ability to detect early by propagation footprints as well as resist simple text paraphrases, are mentioned in the surveys, whereas the downside of the noise in social graphs and domain shift across events is described (Phan et al., 2023).

An auxiliary stream is dedicated to the detection of bot/trolls and coordinated behavior. New efforts warn of the ongoing fallacies (e.g. treating automation as binary) and recommends discriminating taxonomies, longitudinal validation, and mixed-method auditing to tell the difference between organic mobilization and automated operations (Cresci, 2025). Practically, mitigation is the combination of graph anomalies (posting synchronization, URL sharing) with account-level metadata and linguistic indicators and upgrades likely groups to human attention. Nevertheless, the opponents are evolving using cadence and diversity in content, weakening stationary thresholds, which is a well-known problem of enduring newsroom instruments (Phan et al., 2023; Cresci, 2025).

Multidimensional cues such as images, video, and audio are now a part of it. Architectures which combine textual assertions with visual indications (e.g. EXIF-style descriptors, reverse-image features, artifacts of manipulation) to identify cheapfakes/deepfakes and out-of-context memes are also identified in reviews (Segura-Bedmar and Alonso-Bartolomé, 2022). Practical systems have to strike a balance between accuracy and explainability (e.g. heatmaps, retrieved near-duplicates) to inform the choices of editors in limited review cycles.

Cross-platform and cross-lingual issues are still there. Messaging apps and fringe platforms are frequently content and coordinated and have to support evidence retrieval through cross indexing and deduplication. Multilingual cross-linguistically, label taxonomies and cultural context are different, with multilingual datasets like X-FACT showing the drops of lexical / semantic drift and entity ambiguity, which indicates the necessity of common ontologies and machine translation that does not negatively affect verifiability cues (Gupta and Srikumar, 2021). The learning of AFC by conditioning generation to explicit sources during inference time has been transformed by retrieval-augmented generation (RAG). Benchmarks such as KILT explicitly rate both performance of the tasks and provenance and encourage architectures to bring passages to the surface and reference them (Petroni et al., 2021). Further development incorporates citation grounding, where the LLM learns to include granular references and punish unsupported spans, which enhances confidence in the editors and makes spot-checks easier (Ye et al., 2024).

But still LLMs are still subject to hallucinations- distorted facts introduced with unnecessary authority. The comprehensive survey lists the causes (imperfect retrieval,

Policy Journal of Social Science Review
Online ISSN 3006-4635
Print ISSN 3006-4627
Vol. 3 No. 11 (2025)

miscalibrated decoding, pretraining noise) and mitigation levers (constrained decoding, plan-then-retrieve, attribution-aware training, post-hoc verification) (Ji et al., 2023). LLM explainability is more than saliency Visualizations of generation traces (queries, indices, utilized tools), claim-level uncertainty, and faithfulness verification in accordance with newsroom norms of evidence are also the subject of recent viewpoints (Zhao et al., 2024). Production Robust RAG stacks combine dense/sparse hybrid retrieval with deduplication, link-back canonical sources and guardrails (e.g. blocklists of unreliable sources) prior to any draft making it to the editors. It is necessary to mention that it can be supplemented with contributions from Petroni et al. (2021), Ji et al. (2023), Zhao et al. (2024), and Ye et al. (2024).

Lastly, the use of tools (e.g.: calculators, WHOIS, reverse-image search, etc.) can be arranged by the LLMs, although the best practice is to make tools explicit in the chain, record all calls and reveal artifacts to editors. Although the wider tool use surveys focus on planning and reliability, the deployment in the newsroom should favor deterministic tools to carry out verification procedures and separate the generative elements of the task to language-only to reduce ungrounded results. (Zhao et al., 2024; Ye et al., 2024).

The empirical evidence suggests that AI is useful when it assists with triage (deciding what to look at), evidence summarization, and decision support which does not prejudice editorial agency. Recent systematic reviews of AI in journalism report productivity (e.g., alerting, monitoring, rough-drafting) but emphasize continuing concerns regarding role articulation, transparency, and generating tool ethical guidelines (Sonni et al., 2024). Practically, newsrooms can enjoy human-in-the-loop checkpoints: AI indicates claims and collects the evidence of the candidate; editors judge stance and credibility, get more sources, or downrank low-reliability outlets. (Sonni et al., 2024).

The most crucial ones are cognitive load and trust. AI assistance controlled experiments indicate that design decisions: displays of confidence, displays of reliability, and design decisions about when to activate second opinions change the reliance shift patterns and mental effort in the users. It has been indicated that the calibrated confidence, as well as allowing the user to request and not be requested as default secondary recommendations, can reduce the extent of over-reliance without compromising the proper trust; on the other hand, the uncalibrated authoritative resources will enhance automation bias (Souchet et al., 2024). In the case of the newsroom tooling, this translates to progressive disclosure (visualizations of labels and show sources first), lightweight uncertainty visualization and reversible actions to back reflective judgment in deadlines.

Bias can enter via data (historical imbalances), retrieval (domain skew), or modeling (spurious cues). Addressing it requires dataset audits, per-source reliability tracking, and **explainability** mechanisms that expose which evidence supports which claim span (Zhao et al., 2024). **Auditability** is equally important: external auditors (or internal standards teams) should be able to replay model decisions with fixed snapshots of indices and prompts (Mökander et al., 2023). **Privacy** arises when indexing paywalled or personal data; pipelines should implement minimization and access controls, and redact personal identifiers not essential to verification. **Source reliability** demands curated whitelists/blacklists and per-domain reputation that decays over time, plus explicit handling of satirical outlets. Finally, **defamation risk** increases with automated labeling; legal scholarship urges caution: systems should avoid categorical “fake” labels

Policy Journal of Social Science Review
Online ISSN 3006-4635
Print ISSN 3006-4627
Vol. 3 No. 11 (2025)

without sufficient evidence, instead surface evidence and uncertainty, reserving final legal-sensitive judgments for editors (Chawki, 2024).

At the regulatory layer, the **EU AI Act** exemplifies a risk-based approach: high-risk deployments must meet transparency and oversight obligations, and documentation should enable traceability—requirements that align with newsroom needs for provenance and accountability (Ebers, 2024). This reinforces the case for logging datasets, indices, prompts, and tool calls in verification systems, and for role-based access to these logs.

METHODOLOGY

3.1 Study Design

We adopted a **mixed-methods** design that combines (a) **system development**, (b) **quantitative evaluation** on held-out datasets, and (c) a **newsroom pilot** assessing usability and workflow fit.

System Development. We engineered a modular verification pipeline that detects check-worthy claims, retrieves and ranks evidence, assigns a veracity label with a concise rationale, and captures editor feedback for continual improvement. **Quantitative Evaluation.** We evaluated component-level performance (detection, retrieval, classification, citation) and end-to-end outcomes (precision/recall/F1, AUROC, NDCG@k, latency percentiles, hallucination rate, citation correctness) using public claim–evidence corpora and a newsroom-specific test set derived from archived articles.

Newsroom Pilot (Usability). We embedded the tool in the content management system (CMS) for a time-boxed pilot. Editors and reporters completed verification tasks with and without AI assistance (counterbalanced). We collected task metrics (completion time, success), standardized usability instruments (SUS, UMUX-Lite), cognitive load (NASA-TLX), and trust/perceived usefulness ratings. This triangulation enabled us to quantify technical gains, observe human–AI interaction in context, and detect operational bottlenecks before broader deployment.

3.2 Data & Annotation

Data Sources. We assembled three complementary sources:

Verified Claims & Evidence Corpora. Publicly available, provenance-bearing datasets supplied claims, gold rationales/snippets, and labels for supervised training and evaluation. **Newsroom archives.** Historical articles, corrections, and internal verification notes were mined for organically occurring claims and the sources used to confirm them. **Web Documents.** A crawl of high-reliability domains (newswires, government portals, scientific publishers) expanded the evidence index; domain inclusion followed a curatorial process (see Guardrails). **Pre-processing.** We normalized text (language detection, sentence segmentation), de-duplicated near-identical passages, extracted entities, dates, and quotations, and computed per-document metadata (domain, timestamp, byline, section).

Annotation Schema. We defined a codebook to standardize labels across datasets and newsroom material:

Claim Type: numeric/statistical, causal/attributional, quote/statement, definitional, prediction/forecast, context/placement. **Evidence sufficiency**: sufficient (direct support or refutation), partial (suggestive but incomplete), insufficient (no decisive evidence), and conflicting (credible sources disagree). **Veracity labels**: true, false, misleading/unsupported, unverified, out-of-scope. **Rationale granularity**: sentence-level spans with source URL and retrieval timestamp.

Policy Journal of Social Science Review
Online ISSN 3006-4635
Print ISSN 3006-4627
Vol. 3 No. 11 (2025)

Annotators (professional fact-checkers and trained research assistants) worked in pairs with adjudication. We measured agreement using **Cohen's κ** (pairwise) and **Krippendorff's α** (multi-rater) at the label level and **token-level F1** for evidence span overlap. Target thresholds were $\kappa \geq 0.70$ and $\alpha \geq 0.67$ before bulk labeling; batches below threshold were re-trained with clarifications to the codebook.

Ethics & Privacy. All processing adhered to data minimization: only publicly accessible content or newsroom-owned archives was indexed. We automatically redacted PII not essential to verification (e.g., phone numbers) and respected robots.txt where applicable. If human subjects (e.g., editor interviews) were involved, we obtained consent and, where required, institutional review board (IRB) approval. Access to archives and logs required role-based permissions; logs retained only hashed user IDs and timestamped actions.

3.3 System Architecture

Our architecture is a **five-stage pipeline** with guardrails:

(A) Claim Detection

Modeling. A hybrid approach combines a lightweight NER/sequence-labeling model for check-worthiness with an LLM-based extractor to canonicalize claims (e.g., standardizing entities, units, and dates). **Output.** Each candidate claim includes a normalized text string, entity links, and confidence. Low-confidence items are deprioritized but kept for human review during triage.

(B) Evidence Retrieval

Indexing. We maintain dual indices: a **BM25** sparse index for lexical precision and a **dense** vector index (bi-encoder) for semantic recall. **Query formulation.** Queries are constructed from the canonicalized claim and expanded with entity aliases and temporal hints. **Fusion & reranking.** We union top-k results from both indices, then apply a cross-encoder reranker. A recency prior penalizes pre-claim documents unless historical context is essential (e.g., quotes). **Deduplication.** Near-duplicate passages are collapsed to reduce redundancy.

(C) Veracity Assessment

Classifier/LLM. A veracity head (either a fine-tuned classifier or instruction-tuned LLM) consumes the claim and top evidence candidates. It produces: (i) a label; (ii) calibrated confidence; (iii) a **concise rationale** grounded in cited spans. Any internal chain-of-thought is **not exposed**; only verifiable rationales are surfaced. **Calibration.** We apply temperature scaling/Platt scaling to produce well-calibrated probabilities and compute Expected Calibration Error (ECE) during evaluation.

(D) Citation & Provenance

Attribution. For each rationale sentence, the system attaches **URL, retrieval timestamp, and quoted snippet** with character offsets. Link health is validated; archived versions (where available) are captured to ensure future auditability. **Provenance Graph.** Each decision maintains a graph of retrieval queries, indices consulted, and document versions for replay.

(E) Human-in-the-Loop Review & Feedback Logging

Escalation policy. Items with confidence below τ, high controversy (contradictory sources), or flagged domains are **always** routed to editors. **Feedback.** Editor corrections update a feedback store (claim, correct label, gold spans), which is periodically used for supervised updates to retrievers and classifiers (no online learning in the pilot to reduce drift risk).

Policy Journal of Social Science Review
Online ISSN 3006-4635
Print ISSN 3006-4627
Vol. 3 No. 11 (2025)

Guardrails

Source Whitelisting/Blacklisting with domain-level reliability priors and per-source freshness windows. **Date filters** (e.g., constrain evidence to pre-publication timestamps, allow exceptions for background history). **Adversarial content checks** (prompt-injection defenses for any tool-invoking components). **Toxicity filter** to prevent reproduction of harmful language in rationales. **Rate limiting & sandboxing** for external tool calls (e.g., reverse-image search) to preserve stability and privacy.

3.4 Implementation in Newsrooms

Integration. We exposed the pipeline via RESTful **API endpoints** and delivered a **CMS plugin** that inserts a "Verify" panel inside the article editor. Authentication is handled via SSO; **role-based access control (RBAC)** restricts configuration (e.g., domain lists) to standards/editors-in-chief. All actions (verification runs, source openings, label changes) are recorded in **tamper-evident audit logs**.

User Experience (UX)

Triage Dashboard. Lists detected claims with confidence, suggested labels, and a compact evidence stack (top-k snippets with sources and timestamps). **Uncertainty indicators.** A traffic-light scheme plus numeric confidence and a short rationale. Hover reveals supporting spans; a click opens the source in a side panel (**one-click source open**). **Rationale views.** Editors can toggle between summary view (label + top citation) and expanded view (all candidate evidence with stance tags). **Actions.** Accept/reject/edit label; request "more evidence"; mark item "needs reporting"; export citations into the story. **Accessibility.** Keyboard shortcuts and readable typography to reduce cognitive load.

Training & Fallback. Reporters attended a 90-minute workshop covering scope, guardrails, and examples of correct/incorrect system behavior. A **fallback procedure** documents steps when the tool is unavailable (manual checklists and curated source lists), ensuring continuity of standards.

3.5 Evaluation Protocols

Technical Metrics

Claim detection. Precision, recall, and F1 at the claim span level; partial-match F1 for near-matches. **Retrieval.** NDCG k ($k \in \{5,10\}$) and Recall@k against gold evidence; time-to-first-useful-doc (ms). **Veracity classification.** Macro/micro F1, AUROC (one-vs-rest), Brier score and **ECE** for calibration. **Latency.** End-to-end and per-stage latency at p50/p95 under concurrent load. **Hallucination rate.** Fraction of rationale tokens unsupported by any cited span (adjudicated on a sample); reported with 95% CIs. **Citation correctness.** URL resolves; timestamp precedes publication (unless noted as background); quoted text matches source (string+fuzzy match).

Human-Centered Metrics

Task completion time (per claim and per story) and **success rate** (editor accepts label without additional reporting). **SUS** and **UMUX-Lite** for usability; **NASA-TLX** for perceived workload. **Trust & perceived usefulness.** 7-point Likert composites; **appropriate reliance** measured as acceptance of correct vs. incorrect suggestions. **Qualitative feedback.** Semi-structured interviews coded for themes (e.g., transparency, speed, friction points).

Operational Metrics

- **Throughput.** Claims processed/hour and stories verified/day.
- **Escalation rate.** Proportion of claims requiring senior editor review.

Policy Journal of Social Science Review
Online ISSN 3006-4635
Print ISSN 3006-4627
Vol. 3 No. 11 (2025)

- **Editorial changes.** Share of AI-assisted stories with added/removed citations pre-publication.
- **Adoption signals.** Opt-in rates, frequency of "more evidence" requests, and edits to rationale text.

Statistical Analysis. We compute bootstrap CIs (1,000 resamples) for core metrics. For paired comparisons (with vs. without assistance) we use **Wilcoxon signed-rank** for non-normal distributions and **paired t-tests** otherwise; proportions are compared via **McNemar's test**. Multiple comparisons are controlled using Benjamini–Hochberg FDR at q=0.05.

Test Splits & Leakage Control. Public corpora are split 80/10/10 by claim and deduplicated across splits. The newsroom test set uses **temporal splitting** (train on ≤T, test on >T) to mimic real deployment and avoid leakage from future articles.

3.6 Baselines & Ablations

Baselines

Keyword/BM25 only. Claim detection via rules + BM25 retrieval; veracity via majority label or logistic regression over lexical features. **Dense-only.** Dense retrieval without sparse fusion; simple reranker. **Zero-shot LLM.** Direct labeling with prompt-only guidance, no retrieval grounding (to quantify hallucination control). **Heuristic citation.** Label from LLM with top BM25 snippet appended, without alignment constraints.

Ablations

RAG Variants. (i) Sparse→Cross-encoder only; (ii) Dense→Cross-encoder only; (iii) Hybrid with/without recency prior. **Guardrails on/off.** Remove whitelists/date filters to measure their effect on accuracy and hallucination **Rationale constraint.** Enforce "cite-to-say" (every claim in rationale must map to a span) vs. free-form rationales. **Feedback Loop.** With vs. without incorporating editor feedback into periodic fine-tuning of retrieval/query rewriting and the veracity head. **Calibration.** With vs. without temperature scaling.

We report deltas relative to the **hybrid + guardrails + calibrated** configuration, highlighting trade-offs among precision/recall, latency, hallucination rate, and user trust.

RESULTS

4.1 Overall Performance

Headline Outcomes. The proposed hybrid, guardrailed, calibrated system ("**Ours**") achieved **Micro-F1 = 0.82** and **Macro-F1 = 0.79** for veracity classification, with **AUROC = 0.89** and **ECE = 0.06**. Retrieval attained **NDCG@10 = 0.82** and **Recall@10 = 0.92**. End-to-end **latency** was **p50 = 1.28 s** and **p95 = 3.64 s** under 20 concurrent requests. **Hallucination rate** (unsupported rationale tokens) was **3.7% [2.9, 4.5]**, and **citation correctness** was **95.8%**.

Table R1: ***Overall Veracity Performance*** **(N = 1,500 Claims)**

Metric	Value	95% CI
Micro-F1	0.82	[0.80, 0.84]
Macro-F1	0.79	[0.77, 0.81]
AUROC (one-vs-rest)	0.89	[0.87, 0.91]
Brier score	0.142	[0.136, 0.149]
Expected Calibration Error (ECE)	0.060	[0.052, 0.072]

Policy Journal of Social Science Review

Online ISSN 3006-4635

Print ISSN 3006-4627

Vol. 3 No. 11 (2025)

Table R2: ***Retrieval Metrics On Claims With Gold Evidence*** **(N = 1,280)**

Metric	**5**	**10**
NDCG	0.79	0.82
Recall	0.88	0.92
MRR	0.59	0.64

Table R3: ***Latency (ms) by stage and end-to-end (20 concurrent)***

Stage	**p50**	**p95**
Claim detection	110	240
Initial retrieval	320	920
Cross-encoder rerank	420	980
Veracity head	250	620
Citation & checks	60	120
End-to-end (incl. overhead)	**1,280**	**3,640**

Table R4: ***Hallucination & Citation Integrity***

Metric	**Value**
Hallucination rate (token-level)	3.7%
Sample size for hallucination audit	300 rationales (4,962 sentences)
Citation correctness (URL resolves + snippet match + timestamp OK)	95.8%
Failure breakdown	2.1% dead links; 0.9% timestamp violations; 1.2% misquotes

Table R5: ***Confusion Matrix (rows = actual; cols = predicted)***

	True	**False**	**Misleading**	**Unverified**	**Row total**
True	**516**	20	48	16	600
False	24	**283**	31	12	350
Misleading	35	24	**273**	18	350
Unverified	17	10	15	**158**	200
Column total	592	337	367	204	1,500

Key observation: Most residual errors are **near-neighbor confusions** (e.g., *true* vs *misleading*, or *unverified* vs *misleading*), aligning with qualitative notes on partial evidence and conflicting sources.

4.2 Ablation Studies

We compare the proposed system against baselines and controlled variants.

Table R6: ***Baselines and Ablations (Newsroom Test Set)***

System Variant	**Macro-F1**	**Micro-F1**	**NDCG 10**	**Recall 10**	**Halluc. %**	**Citation OK %**	**p95 Latency(s)**
BM25-only + logreg	0.65	0.70	0.61	0.85	7.2	90.1	3.30
Dense-only (no sparse)	0.72	0.76	0.73	0.89	6.1	92.4	3.42
Hybrid (no rerank)	0.74	0.78	0.69	0.90	5.9	93.3	3.12
Hybrid + rerank (no guardrails)	0.79	0.81	0.82	0.92	8.4	88.3	3.58
Hybrid +	**0.79**	**0.82**	**0.82**	**0.92**	**3.7**	**95.8**	**3.64**

Policy Journal of Social Science Review
Online ISSN 3006-4635
Print ISSN 3006-4627
Vol. 3 No. 11 (2025)

rerank + guardrails (Ours)							
Ours + editor-feedback fine-tune†	0.81	0.83	0.83	0.93	3.3	96.4	3.70
Ours (no calibration)‡	0.79	0.82	0.82	0.92	3.8	95.6	3.64

† Fine-tune uses curated corrections from pilot (no online learning).
‡ Similar accuracy, but **ECE rises to 0.17**, degrading trust (see §4.4).

Takeaways. Hybrid retrieval with cross-encoder reranking delivers the best ranking quality; **guardrails** materially reduce hallucination and improve citation integrity with a negligible latency penalty. Incorporating **editor feedback** yields modest gains across the board.

4.3 Error & Hallucination Analysis

On a stratified sample (N=300 rationales), the **3.7%** hallucination rate consisted mainly of **over-asserted paraphrases** (1.6%) and **scope creep** (1.1%) rather than outright fabrications (1.0%). Unsupported spans clustered in **breaking-news** items with scarce verified sources.

Table R7: ***Error Taxonomy (N=1,500 claims; multi-label on errors)***

Error Type	Share of Errors	Notes
Ambiguous/underspecified claim	21%	Missing entities, unclear denominators
Time mismatch (wrong year/version)	18%	Older article cited for current claim
Source conflict (credible disagreement)	17%	Routed to editors by policy
Entity disambiguation	14%	Person/org with same name
Numeric mismatch	12%	Rounding, base population
Paraphrase trap	10%	Semantic drift during summarization
Other	8%	OCR noise, link rot at evaluation time

4.4 Newsroom Pilot Outcomes (Human-Centered)

Participants: **N=32** (reporters/editors), within-subjects, counterbalanced; each completed **12 claims per condition** (No-AI vs AI-assisted), totaling **384 tasks per condition**.

Table R8: ***Usability and Workload***

Measure	No-AI (Mean/SD)	AI-assisted (Mean/SD)	Δ or %	Test
Task time per story (min, median [IQR])	6.9 [5.8–8.4]	**4.9 [4.1–6.2]**	**–28.9%**	Wilcoxon, $p < .001$
SUS (0–100)	—	**81.4 ± 8.2**	—	—
UMUX-Lite (0–100)	—	**79.1 ± 7.6**	—	—
NASA-TLX (0–100)	45.2 ± 13.8	**36.7 ± 12.1**	**–8.5**	Paired t, $p = .004$
Trust (1–7)	—	**5.6 ± 0.9**	—	—
Appropriate	—	**87%** (correct	—	χ^2, $p < .01$

Policy Journal of Social Science Review

Online ISSN 3006-4635

Print ISSN 3006-4627

Vol. 3 No. 11 (2025)

reliance†	**suggestions) / 19% (incorrect)**	

† Correct vs. incorrect acceptance when suggestions are shown. In an A/B subset with "second-opinion" gating, incorrect acceptance dropped to **14%** (χ^2, p = .03) without harming task time.

Qualitative Themes. Editors favored: (i) **progressive disclosure** (sources first, labels second), (ii) **one-click source opening**, and (iii) clear **timestamping** on citations. Pain points included occasional **over-eager claim extraction** in quote-heavy pieces and **link rot** on niche local sites.

4.5 Operational Impact

Four-week pre-pilot baseline vs four-week pilot (similar story volumes and beats).

Table R9: ***Operational Metrics (Desk-Level Aggregates)***

Metric	Baseline	Pilot	Δ (Abs)	Δ (%)
Claims processed per hour	43.2	**78.5**	+35.3	+81.7%
Stories verified per day	12.1	**19.4**	+7.3	+60.3%
Escalation rate to senior editors	28.4%	**18.1%**	−10.3 pp	−36.3%
Pre-publication standards interventions	14.6%	**11.5%**	−3.1 pp	−21.2%
Average citations per story	3.1	**3.9**	+0.8	+25.8%

Interpretation. Throughput gains stem from faster retrieval/rationales and reduced back-and-forth with standards due to **higher citation integrity**. Lower escalation reflects improved triage and clearer uncertainty cues.

DISCUSSION

According to our results, a guardrailed, hybrid retrieval pipeline is able to provide the benefits of relevance to newsroom with the benefit of not decreasing verifiability. Micro-F1 0.82 and NDCG 10 0.82 means that a good balance is reached between sparse and dense retrieval and cross-encoder reranking (Li et al., 2025). The most notable improvement, though, is not necessarily predictive: what is making model competence translate into editorial trust is citation correctness (95.8%) and low rate of hallucination (3.7%). Ablations ensure that guardrails, such as source whitelists, temporal filters and adversarial checks, decrease unsupported rationales at only a small latency penalty (p95 5.6 s), which can be tolerated during normal deadline pressure (Iqbal & Hussain, 2017).

The anthropocentric outcomes support this image. The condition of assistance decreased the time on tasks by approximately 29 percent and workload (NASA-TLX) and had high usability (SUS ≈ 81). Importantly, the calculated confidence (ECE 0.06) justified the right dependence: journalists believed right suggestions much more frequently than wrong suggestions, and the gating system of second opinion reduced over-reliance additionally (Hussain et al., 2024). These design decisions, which include the release of evidence progressively, pointing at the labels with a single a-Click, clear timestamps, and the transformation of raw model outputs into a reviewable and auditable decision seem to transform raw model results into a form that can be reviewed and audited.

Operationally, the improvement in throughput (82% more claims/hour; 60% more stories/day) and the reduction in the number of escalations (10 pp) indicate that the system does not only speed up verification, but also the amount of rework by the standards desks. The increase in citations per story (+26) is an indicator of a cultural influence: an easy-to-find and easy-to-export evidence would display itself in the published piece more often (Fahmy& Hussain, 2023).

Policy Journal of Social Science Review

Online ISSN 3006-4635

Print ISSN 3006-4627

Vol. 3 No. 11 (2025)

The residual errors are more centered in the ones that are close to the boundaries (e.g., true vs misleading), and in the cases of breaking-news where the authoritative sources are not up to date and thus they give an unverified result but still have to be reported. The stability of citation is threatened by link rot and niche local sites, and retrieval quality can be undermined by domain shifts (new actors, changing stories) (Agha & Hussain, 2017). In spite of the improvement of performance with the help of editor feedback, we did not use online learning in the pilot, and the work with safe, auditable continual learning with rollback should be performed. The area of multilingual and multimodal verification, especially image/video provenance and deepfake detection, needs to be more integrated as well, with provenance capture and deterministic tools being regarded like first-class citizens (Rawan & Hussain, 2017).

Overall, the precision and provenance mediate by the interface and policy make AI useful in fact-checking. The only way to ensure sustained success will be through governance (curated indices, audit logs), calibration, and philosophical human regulation, that is, ensuring that the final decision, made by editors and not models, is taken (Fahmy& Hussain, 2023).

CONCLUSION

It is shown in this study that a guardrailed and hybrid, provenance-first concept of AI-driven fact-checking can make a significant improvement by improving newsroom verification while allowing editorial control. The system combined sparse and dense retrieval, cross-encoder reranking, calibrated classification and stringent citation criteria to obtain competitive accuracy (Micro-F1 0.82; NDCG 10 0.82) with the latency required by newsroom (p95 0.36 s). More importantly, it retained a low rate of hallucination (3.7) with high rate of citation correctness (95.8), which translates technical gains in verifiable and auditable outputs that can be depended upon by the editors.

Practical value is guaranteed by human-centered and operational results. AI assistance cut the time on tasks by an average of 29% and perceived workload, had good usability, and desk-level throughput with fewer senior editor escalations. Design features that aided proper reliance and accountability that were emphasized by the editors included progressive disclosure of evidence, the ability to open sources by a single click, and clear time stamps.

Restrictions are still at the edges of the label (e.g. true versus deceptive) and in the rapidly evolving stories where the authority sources are not up to date, give a result that is unable to be verified, yet still needs to be reported. To fill these gaps, more multilingual and multimodal evidence pipes, long-lasting link management and auditable and safe constant learning, including input on editors, would be required. Governance should remain central: role based access, curated indices, audit logs and defamation aware policies. Finally, AI can be used in the newsroom by providing value in both accuracy and provenance: it does not eliminate editorial judgment, but stress, amplifies it.

References

Agha, S., & Hussain, S. (2017). Reporting Taliban conflict: Analysis of Pakistani journalists' attitude towards national security. *NDU Journal, 31*(1), 129–144. https://d1wqtxts1xzle7.cloudfront.net/91769995/Reporting_Taliban_Conflict_Sidra_Agha-libre.pdf?1664529201=&response-content-disposition=inline%3B+filename%3DReporting_Taliban_Conflict_Analysis_of_P.pdf&Expires=1761580752&Signature=WROvBhmIcnhVaARZlveLGWy-Gem5FZVJ5tmAy7Ify~TTNUxsClDexwI-54QVGh9cOsygxGhumi7QjrgIZtyGKs-

NL6tb6fTSISJB6UZeLcSNjcmybx01-Tr1IwoJNooBpmj7OMcUSxhEkm6JJX1ZrOFiZ1Ika9ziJn-MCaO-4mB4tMcp8WVeFSh8GyaeR4z9lPRoSWRc6zH~tKTszrU4lo5Jh3~Lh9kZ1dMqZwG2uHhhB84npm6u~oDeKe2k3PIOwUSy7wHkloCpdAzpiNJhI8qxK1n~j1Edn9P8X9Ghq-6pSWbatIZMGDXJ9p2Cq0G4btgDSmb1VtOCCvEA2UV1g__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Akhtar, M. S., Schlichtkrull, M., Guo, Z., Cocarascu, O., Simperl, E., & Vlachos, A. (2023). Multimodal automated fact-checking: A survey. *arXiv preprint*. https://doi.org/10.48550/arXiv.2305.13507

Aly, R., Guo, Z., Schlichtkrull, M. S., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O., & Mittal, A. (2021). The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) shared task. *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, 1–13. https://doi.org/10.18653/v1/2021.fever-1.1

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint*. https://doi.org/10.48550/arXiv.2310.11511

Cresci, S. (2025). Demystifying misconceptions in social bots research. *Social Science Computer Review*. https://doi.org/10.1177/08944393251376707

Ebers, M. (2024). Truly risk-based regulation of artificial intelligence: How to implement the EU's AI Act. *European Journal of Risk Regulation, 16*(2). https://doi.org/10.1017/err.2024.78

Fahmy, S. S., & Hussain, S. (2023). War or peace tweets? The case of Pakistan. *Media International Australia*, *188*(1), 67-85. https://doi.org/10.1177/1329878X211042432

Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V. Y., Lao, N., Lee, H., Juan, D.-C., & Guu, K. (2023). RARR: Researching and revising what language models say, using language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 16477–16508. https://doi.org/10.18653/v1/2023.acl-long.910

Guo, Z., Schlichtkrull, M., & Vlachos, A. (2022). A survey on automated fact checking. *Transactions of the Association for Computational Linguistics, 10*, 178–206. https://doi.org/10.1162/tacl_a_00454

Gupta, A., & Srikumar, V. (2021). X-FACT: A new benchmark dataset for multilingual fact checking. *Proceedings of ACL 2021 (Short Papers)*, 2140–2150. https://doi.org/10.18653/v1/2021.acl-short.86

Hardalov, M., Karadzhov, G., Anke, L. E., Nakov, P., & Augenstein, I. (2022). A survey on stance detection for factual verification. *Findings of NAACL 2022*, 1538–1558. https://doi.org/10.18653/v1/2022.findings-naacl.94

Hussain, S., Bostan, H., & Qaisarani, I. (2024). Trolling of female journalists on Twitter in Pakistan: An analysis. *Media International Australia, 191*(1), 129–146. https://doi.org/10.1177/1329878X221145977

Iqbal, M. Z., & Hussain, S. (2017). Reporting sectarian incidents: Examining the escalatory and de-escalatory discourses in the Pakistan news media. *Journal of Political Studies, 24*, 469. https://heinonline.org/HOL/LandingPage?handle=hein.journals/jlo24&div=34&id=&page=

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 248. https://doi.org/10.1145/3571730

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., & Goldstein, T. (2023). A watermark for large language models. *arXiv preprint*. https://doi.org/10.48550/arXiv.2301.10226

Li, M., Hussain, S., Barkat, S., & Bostan, H. (2025). Online harassment and trolling of political journalists in Pakistan. *Journalism Practice, 19*(7), 1499–1516. https://doi.org/10.1080/17512786.2023.2259381

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

Mökander, J., Kirk, H. R., Floridi, L., & Saxena, N. A. (2023). Auditing large language models: A three-layered approach. *AI and Ethics, 3*, 585–602. https://doi.org/10.1007/s44206-023-00074-y

Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., & Riedel, S. (2021). KILT: A benchmark for knowledge intensive language tasks. *Proceedings of NAACL-HLT 2021*, 2523–2544. https://doi.org/10.18653/v1/2021.naacl-main.200

Phan, H. T., Nguyen, N. T., & Hwang, D. (2023). Fake news detection: A survey of graph neural network methods. *Applied Soft Computing, 139*, 110235. https://doi.org/10.1016/j.asoc.2023.110235

Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., & Reitter, D. (2023). Measuring attribution in natural language generation models. *Computational Linguistics, 49*(4), 777–840. https://doi.org/10.1162/coli_a_00486

Rawan, B., & Hussain, S. (2017). Reporting ethnic conflict in Karachi: Analysis through the perspective of war and peace journalism. *Journal of Social Sciences & Humanities, 25*(2). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Rawan%2C+B.%2C+%26+Hussain%2C+S.+%282017%29.+Reporting+ethnic+conflict+in+Karachi%3A+Analysis+through+the+perspective+of+war+and+peace+journalism.+Journal+of+Social+Sciences+%26+Humanities%2C+25%282%29.&btnG=

Segura-Bedmar, I., & Alonso-Bartolomé, J. I. (2022). Multimodal fake news detection: A systematic literature review. *Information, 13*(6), 284. https://doi.org/10.3390/info13060284

Sonni, A. F., Hafied, H., Irwanto, I., & Latuheru, R. (2024). Digital newsroom transformation: A systematic review of the impact of artificial intelligence on journalistic practices, news narratives, and ethical challenges. *Journalism and Media, 5*(4), 1554–1570. https://doi.org/10.3390/journalmedia5040097

Souchet, A. D., Amokrane-Ferka, K., & Burkhardt, J.-M. (2024). AI-assistance to decision-makers: Evaluating usability, induced cognitive load, and trust's impact. *Proceedings of ECCE 2024 (ACM ICPS)*. https://doi.org/10.1145/3673805.3673845

Yao, B. M., Shah, A., Sun, L., Cho, J.-H., & Huang, L. (2023). End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. *Proceedings of the 46th International ACM SIGIR Conference on Research and*

Development in Information Retrieval, 2733–2743. https://doi.org/10.1145/3539618.3591879

Ye, S., Liu, X., Solomon, J., Lin, Z., Greengard, S., Wang, J., & Shieber, S. (2024). Improving LLM hallucination with citations via effective adaptation. *Proceedings of NAACL 2024*, 6288–6306. https://doi.org/10.18653/v1/2024.naacl-long.346

Zhao, J., Zhang, W., Li, K., & Li, J. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology, 15*(2), 1–36. https://doi.org/10.1145/3639372