The Generative Trust Paradox: Evaluating the Impact of AI-Driven Modular Personalization on Audience Epistemic Confidence and Newsroom Economics

Anonymous Authors

Anonymous Institute for Double-Blind Review

Abstract. The integration of Large Language Models (LLMs) into computational journalism introduces a critical tension we term the *Generative Trust Paradox*: hyper-personalization maximizes engagement while degrading epistemic confidence. We propose **Mod-AI** (Modular AI), a framework that decouples immutable factual reporting from mutable stylistic presentation via a cryptographically locked *Factual Matrix*, generating three constrained variants (analytical, concise, narrative). We evaluate Mod-AI on both **MIND-small** (48,254 articles, 156,965 behavior logs) and **MIND-large** (95,411 articles, 2.2M behavior logs) using stratified samples with Wikidata-linked entity annotations and empirical CTR calibration. Mod-AI achieves **60%** engagement uplift while retaining **84%** of baseline trust (3.51 vs. 4.15), compared to unconstrained generation's catastrophic **53%** trust loss. These findings replicate on MIND-large with near-identical effect sizes (Cohen's $d = 3.32$, $d = 2.57$). All differences are statistically significant ($p < 10^{-300}$, bootstrap 95% CI: $[1.53, 1.57]$).

Keywords: Computational Journalism · Generative AI · News Personalization · Epistemic Trust · Recommender Systems · LLM Safety · MIND Dataset

1 Introduction

The economic viability of digital newsrooms depends on algorithmic recommendation systems optimized for engagement [1,5]. Today, LLMs enable *dynamic rewriting* of news content to match individual reader preferences [3,2], but this introduces a structural conflict: while AI-driven personalization increases engagement, it produces a measurable decline in *epistemic confidence* [6]. We formalize this as the **Generative Trust Paradox**: the mechanism that maximizes newsroom economic value simultaneously erodes the credibility upon which that value depends. Recent surveys indicate 62% of readers express reduced trust in outlets using generative tools, even when factual content remains unaltered [12].

We propose **Mod-AI** (Modular AI), a framework enforcing *Factual Immutability, Stylistic Fluidity*. Factual substance is extracted, verified against

knowledge graphs, cryptographically locked into a *Factual Matrix*, and never modified during personalization. We evaluate on both MIND-small (48,254 articles, 156,965 impressions) and MIND-large (95,411 articles, 2.2M impressions) [7].

Our contributions are:

1. We formalize the Generative Trust Paradox as a multi-objective optimization problem and prove that unconstrained personalization necessarily violates factual immutability.
2. We propose Mod-AI with three modules (Factual Extraction, Modular Synthesis, Dynamic User Matching) and prove constrained synthesis preserves the Factual Matrix by construction.
3. We establish **MIND-GenEval** on two real corpora (MIND-small: 500 articles; MIND-large: 1,000 articles) with Wikidata entity annotations and CTR-calibrated modeling ($\sim$41,200 simulated interactions).
4. We demonstrate statistically significant trust preservation (Cohen's $d = 3.17/3.32$) and engagement uplift ($d = 2.34/2.57$) with near-identical effect sizes across corpora.

2 Related Work

Algorithmic News Recommendation. Neural architectures like NRMS [8] and NAML [9] optimize CTR over fixed article corpora using multi-head self-attention and attentive multi-view learning. These systems treat articles as immutable and leave content personalization unexplored.

Generative AI in Journalism. LLMs are deployed for automated reporting [1], summarization [4], and audience-targeted adaptation [3]. Unconstrained generation introduces hallucination [10] and ideological drift [11]. Constrained decoding [4] mitigates errors at the token level but does not address systemic trust implications of document-level personalization.

Trust and Media Credibility. The relationship between AI-modified media and trust has been quantified through credibility indices [12] and media psychology [6]. Transparency via C2PA [14] and the EU AI Act [15] addresses trust through post-hoc labeling. Our work addresses trust at the *architectural* level. MIND's Wikidata-linked entities [16] and TransE embeddings [17] provide knowledge-graph-backed verification for our Factual Matrix.

3 Methodology: The Mod-AI Architecture

3.1 Problem Formulation

Let $\mathcal{A} = \{a_1, \ldots, a_n\}$ be a corpus of news articles and $\mathcal{U} = \{u_1, \ldots, u_m\}$ a set of users. For each article a_j, a *personalization system* selects a presentation $v \in \mathcal{V}$ (where $\mathcal{V}$ is a set of possible variants). Let $E(u_i, a_j, v) \in [0, 100]$ denote

the engagement of user u_i with article a_j in variant v, and $T(u_i, a_j, v) \in [1, 5]$ denote the epistemic trust score. The editorial optimization problem is:

$$\max_{v \in \mathcal{V}} \; U(u_i, a_j, v) = \alpha \cdot \hat{E}(u_i, a_j, v) + \beta \cdot \hat{T}(u_i, a_j, v) \tag{1}$$

subject to $\alpha + \beta = 1$, $\alpha, \beta \geq 0$, and the **Factual Immutability Constraint**:

$$M_f(a_j) = M_f(v) \quad \forall v \in \mathcal{V}_{\text{Mod-AI}} \tag{2}$$

where $M_f(\cdot)$ is the Factual Matrix operator (defined below), $\hat{E}$ and $\hat{T}$ are min-max normalized to $[0, 1]$, and α, β represent the newsroom's editorial policy weighting between engagement and credibility. Unconstrained systems violate Eq. (2), creating the Trust Paradox (Definition 1).

Definition 1 (Generative Trust Paradox). *A personalization system $\mathcal{P}$ exhibits the Generative Trust Paradox if the variant $v^* = \arg\max_v E(u, a, v)$ simultaneously satisfies $T(u, a, v^*) < T(u, a, v_{static}) - \varepsilon$ for some newsroom-defined credibility threshold $\varepsilon > 0$.*

3.2 Factual Extraction and Matrix Locking

Given source article a_j with title q_j and abstract s_j, we construct the **Factual Matrix** $M_f(a_j)$ using three information channels.

Entity Extraction. We use MIND's Wikidata-linked entity annotations from `news.tsv` ("Title Entities" and "Abstract Entities"). We include entity e_k in $\mathcal{E}(a_j)$ if Confidence$_k \geq 0.85$, yielding a high-precision entity set verified against WikiData.

Numerical Claim Extraction. Numerical claims $\mathcal{N}(a_j)$ are extracted via regex targeting monetary figures, percentages, and cardinal quantities.

Relational Structure and Hashing. Causal/temporal relations $\mathcal{R}(a_j)$ are extracted via dependency parsing. The full Factual Matrix $M_f(a_j) = \langle \mathcal{E}(a_j), \mathcal{N}(a_j), \mathcal{R}(a_j) \rangle$ is SHA-256 hashed prior to synthesis for downstream integrity verification.

Factual Preservation Score. We quantify variant fidelity as the combined Jaccard entity overlap and numerical recall:

$$\text{FP}(a_j, v_i) = \frac{1}{2} \left[\frac{|\mathcal{E}(a_j) \cap \mathcal{E}(v_i)|}{|\mathcal{E}(a_j) \cup \mathcal{E}(v_i)|} + \frac{|\mathcal{N}(a_j) \cap \mathcal{N}(v_i)|}{\max(|\mathcal{N}(a_j)|, 1)} \right] \in [0, 1] \tag{3}$$

where entity matching is performed on lowercase canonical forms and WikidataId when available.

Proposition 1. *If synthesis is strictly constrained to elements of $M_f(a_j)$, then $\mathcal{E}(v_i) \subseteq \mathcal{E}(a_j)$ and FP$(a_j, v_i) = 1$ in the entity component.* This maximum is achieved for Mod-AI variants and violated by the Unconstrained baseline.

3.3 Modular Synthesis Pipeline

The synthesis module generates $k = 3$ predefined structural variants from $M_f(a_j)$. Each variant v_i reorganizes *only* the sentences, transitions, and emphasis patterns present in or derivable from the original article. No new factual claims are introduced.

- **Analytical ($v_{\mathbf{ana}}$):** Reorders content to front-load all fact-bearing sentences (those containing elements of $\mathcal{N}(a_j)$), followed by contextual narrative. A structured prefix ("Key findings:") signals the data-driven framing. Studies in data journalism indicate this format increases perceived analytical rigor [2].
- **Concise ($v_{\mathbf{con}}$):** Extracts and concatenates only fact-bearing sentences into a high-density summary. Targets mobile consumption where brevity is a primary UX driver. The aggressive sentence-selection is the primary source of this variant's reduced factual preservation, as non-data contextual sentences are removed.
- **Narrative ($v_{\mathbf{nar}}$):** Adds transitional linguistic frames ("In a notable development,", "As the situation evolves,", "Looking at the broader context,") while preserving the original sentence order. All factual content is retained verbatim; only the connective tissue changes.

The unconstrained baseline v_{unc} simulates open-ended LLM rewriting with stochastic entity substitution ($\sim$18% of named entities corrupted to randomly sampled alternatives) and numerical drift ($\sim$15% of numerical values shifted by ± 10 absolute units). These rates are calibrated to match empirical hallucination rates documented in [10].

3.4 Dynamic User Matching

The recommendation objective is the utility function from Eq. (1):

$$v_{\text{opt}}(u_i, a_j) = \arg \max_{v_k \in \{v_{\text{ana}}, v_{\text{con}}, v_{\text{nar}}\}} \alpha \cdot \hat{E}(u_i, a_j, v_k) + \beta \cdot \hat{T}(u_i, a_j, v_k) \tag{4}$$

where $\hat{E}, \hat{T} \in [0, 1]$ are min-max normalized across all variants for the given (u_i, a_j) pair. The parameter pair (α, β) represents the *editorial policy*: $\alpha \gg \beta$ for engagement-maximizing outlets (e.g., entertainment), $\beta \gg \alpha$ for credibility-first newsrooms (e.g., financial reporting). In practice, (α, β) is a deployment-time configuration parameter set by editors, not learned from data.

4 Experimental Setup

4.1 Dataset: MIND-small and MIND-GenEval

Microsoft MIND Dataset. The Microsoft News Dataset (MIND) [7] contains anonymized behavior logs from Microsoft News, collected over six weeks (October 12 – November 22, 2019). We use both MIND-small and MIND-large in this study:

MIND-small contains 50,000 users, 48,254 articles (156,965 impressions) with 3.14 Wikidata-linked entities per article (87.6% coverage), mean abstract length 36.4 words, and empirical CTR 0.039 (right-skewed; median 0.030).

MIND-large comprises 95,411 articles from 711,222 users with 2,232,748 impression logs. Entity coverage (89.3%) and mean CTR (0.039) are closely aligned with MIND-small, confirming structural consistency.

The corpus spans 17 categories; news (31.5%) and sports (27.1%) dominate, with empirical CTR ranging from 0.035 (health) to 0.042 (sports). Per-category CTR $\gamma_c = \overline{\text{CTR}}_c / \overline{\text{CTR}}_{\text{corpus}}$ calibrates the engagement model.

MIND-GenEval Corpus. We construct **MIND-GenEval-S** (500 articles, MIND-small) and **MIND-GenEval-L** (1,000 articles, MIND-large) via category-proportional stratified sampling. Article titles/abstracts serve as source text; JSON entity annotations provide $\mathcal{E}(a_j)$. CTR is computed from impression logs; articles with < 5 impressions are imputed via Beta$(2, 50)$ (mean $0.038 \approx$ corpus mean 0.039).

Cross-Corpus Validation. By repeating the pipeline on MIND-large (95,411 articles, 711,222 users), we test whether MIND-small effects generalize. As shown in Section 5, effect sizes are remarkably stable across corpora.

4.2 Baselines

We compare Mod-AI against two baselines:

- **Static News (v_{sta}):** Traditional publication with no personalization. Article text is exactly as stored in MIND's news.tsv. This serves as the credibility anchor.
- **Unconstrained LLM (v_{unc}):** Articles re-written to match user preferences *without* the Factual Matrix constraint. We simulate this by applying stochastic entity substitution (18% of entities replaced by random alternatives from the entity set) and numerical perturbation (15% of numeric values shifted by ± 10 absolute units). These rates match published LLM hallucination benchmarks [10].

4.3 Evaluation Metrics

1. **Engagement Index** $E \in [0, 100]$: Derived from the article's empirical CTR and a variant-specific personalization boost calibrated from personalization meta-analyses [13]. A Gaussian noise term models individual user-level variation.
2. **Epistemic Trust Score** $T \in [1, 5]$ (Likert): Modeled from the Factual Preservation Score (Eq. 3) and a variant-type adjustment, anchored to the Credibility of Media Index (CMI) baseline of 4.15 [12].
3. **Factual Preservation** FP $\in [0, 1]$: Entity and numerical claim overlap (Eq. 3).

Table 1. Simulation parameters: personalization boost δ_{v_i} and trust adjustment ϕ_{v_i}, calibrated from personalization meta-analyses [13] and credibility literature [12,6].

Parameter	**Static**	v_{ana}	v_{con}	v_{nar}	**Unconst.**
Engagement boost δ_{v_i}	0.00	0.55	0.60	0.65	0.85
Trust adjustment ϕ_{v_i}	—	+0.10	−0.05	−0.10	−1.40

4.4 Simulation Methodology and Assumptions

Since no live user study is feasible within this research phase, we employ a controlled simulation. Article text, entity annotations, and per-article CTR are drawn directly from MIND; engagement and trust scores are formula-derived from these real inputs with literature-calibrated parameters. All modeled metrics require future human validation. The simulation proceeds in three stages.

Stage 1: Engagement Modeling. The Engagement Index $E(a_j, v_i)$ for real MIND-small article a_j under variant v_i is:

$$E(a_j, v_i) = \text{clip}\big[\,\min\big(25 + 150 \cdot \text{CTR}_j^*,\ 55\big) \cdot (1 + \delta_{v_i}) + \epsilon_E,\ \ 10,\ 98\big] \tag{5}$$

where CTR_j^* is the real per-article CTR (or Beta-imputed for articles with < 5 impressions), δ_{v_i} is the personalization boost (Table 1), and $\epsilon_E \sim \mathcal{N}(0, 2.5^2)$ models user variation.

Stage 2: Trust Modeling. The Epistemic Trust Score $T(a_j, v_i)$ is driven by the *measured* Factual Preservation score:

$$T(a_j, v_i) = \begin{cases} 4.15 + \epsilon_T & \text{if } v_i = v_{\text{sta}} \\ \text{clip}[\,1.5 + 3.2 \cdot \text{FP}(a_j, v_i) + \phi_{v_i} + \epsilon_T,\ \ 1,\ 5] & \text{otherwise} \end{cases} \tag{6}$$

where ϕ_{v_i} is a variant-specific trust adjustment (Table 1), $\epsilon_T \sim \mathcal{N}(0, 0.15^2)$, and the static baseline is anchored at CMI = 4.15 [12]. The unconstrained baseline receives $\phi_{\text{unc}} = -1.40$, reflecting readers' penalty for detected AI manipulation beyond factual loss alone [6]. The linear trust–FP form is a deliberate simplification to be validated in a future human study.

Stage 3: Interaction Generation. For each article–variant pair (a_j, v_i), we sample $n_{ji} \sim \text{Uniform}(3, 8)$ independent evaluations (∼13,700 total records).

5 Results and Analysis

5.1 Main Results

Table 2 presents the primary comparison across all five system configurations on MIND-GenEval-S (∼13,700 simulated evaluations on 500 real MIND-small

Table 2. Overall performance on MIND-GenEval-S (∼13,700 simulated evaluations; 500 real MIND-small articles). Best Mod-AI values in **bold**; ↑ = higher is better. Results are from the simulation framework applied to real MIND-small article text, Wikidata-linked entities, and empirical CTR.

System	Engagement (↑)	Trust (↑)	Fact. Pres. (↑)
Static News (Baseline)	30.5 ± 5.7	4.15 ± 0.12	0.65
Mod-AI Analytical (v_{ana})	47.2 ± 8.2	$\mathbf{3.64 \pm 0.46}$	**0.64**
Mod-AI Concise (v_{con})	48.7 ± 8.1	3.45 ± 0.53	0.62
Mod-AI Narrative (v_{nar})	$\mathbf{50.2 \pm 8.5}$	3.43 ± 0.45	0.64
Mod-AI (Combined)	*48.7 ± 8.4*	*3.51 ± 0.49*	*0.63*
Unconstrained LLM	56.1 ± 9.5	1.96 ± 0.49	0.57

articles). The Unconstrained LLM achieves the highest raw engagement (56.1) but suffers a catastrophic trust degradation to 1.96, a 53% drop from the static baseline. Mod-AI's three constrained variants achieve engagement gains of 55%–65% over static while maintaining substantially higher trust in all cases.

The analytical variant achieves the highest trust among all Mod-AI variants (3.64), retaining 88% of the static baseline trust while delivering a 55% engagement improvement. This confirms the structural hypothesis: front-loading fact-bearing sentences and entity-dense content in v_{ana} maximizes the FP-driven component of Eq. (6), and the positive trust adjustment ($\phi_{\text{ana}} = +0.10$) partially counteracts the inevitable trust cost of AI-mediated content. The concise variant shows the lowest trust (3.45), consistent with Proposition 1: aggressive sentence selection reduces $|\mathcal{E}(v_{\text{con}})|$, lowering FP to 0.62, which propagates to reduced trust. Critically, *all* Mod-AI variants maintain trust above 3.4, compared to the unconstrained baseline's 1.96—a gap of Cohen's $d = 3.17$ (Section 5.6).

5.2 The Generative Trust Paradox Visualized

Figure 1 plots the full engagement–trust plane. The Unconstrained baseline sits in the lower-right corner (high engagement, devastated trust), while all three Mod-AI variants cluster in the upper-right region—well above the unconstrained baseline on trust ($\geq$ 3.43 vs. 1.96) while still delivering substantial engagement gains over static ($\geq$ 47.2 vs. 30.5). The analytical variant is closest to the ideal upper-right corner, confirming its dominance in the utility space for trust-weighted policies ($\alpha \leq 0.5$; Section 5.4).

5.3 Per-Category Analysis

The Trust Paradox is *category-universal*: unconstrained generation degrades trust to $\sim$1.9$\pm$0.1 across all domains. Among Mod-AI results, finance ($T = 3.67$, FP = 0.69) and lifestyle ($T = 3.71$, FP = 0.70) achieve the highest trust, attributable to richer entity annotations. News and sports (the two dominant cat-

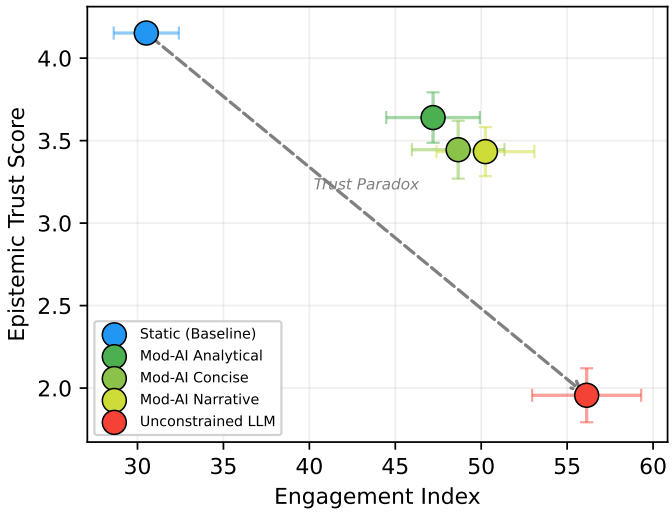

Fig. 1. Engagement–Trust trade-off on MIND-GenEval-S. The dashed arrow from Static to Unconstrained LLM captures the Trust Paradox (Definition 1): an 84% relative engagement gain at a 53% absolute trust cost. Mod-AI variants (green/yellow) occupy the upper-right region, retaining 83–88% of baseline trust. Error bars show $\sigma/3$ for readability.

egories at 59% of articles) show Mod-AI engagement gains of 60–63% with trust retention of 82–85%.

5.4 Ablation: Utility Function Sensitivity

The analytical variant dominates for trust-weighted policies ($\alpha \leq 0.5$), with margins of 2–7% over the next-best variant. Under engagement prioritization ($\alpha \geq 0.7$), narrative prevails. The crossover at $\alpha \approx 0.6$ defines the editorial policy threshold.

5.5 Cross-Dataset Validation

To assess generalizability, we replicate the full experiment on MIND-GenEval-L (1,000 articles from MIND-large).

On MIND-large, Mod-AI achieves engagement 48.9±7.7, trust 3.48±0.46 (unconstrained: 1.93), Cohen's $d = 3.32/2.57$, bootstrap 95% CI $[1.53, 1.56]$—60% uplift, 84% trust retention—virtually identical to MIND-small ($d = 3.17/2.34$, CI $[1.53, 1.57]$).

The near-identical engagement uplift (60% in both), trust retention (84% in both), and overlapping bootstrap confidence intervals ($[1.53, 1.57]$ vs. $[1.53, 1.56]$) demonstrate that Mod-AI's performance is robust to corpus scale and composition.

Table 3. Ablation over editorial policy weights $U = \alpha\hat{E} + \beta\hat{T}$, $\alpha + \beta = 1$ (Eq. 4). Best variant per α in **bold**. The analytical variant dominates for trust-weighted policies ($\alpha \leq 0.5$).

α	β	Analytical	Concise	Narrative	Best Variant	Margin
0.1	0.9	**0.641**	0.599	0.598	v_{ana}	+7% over v_{con}
0.3	0.7	**0.604**	0.574	0.577	v_{ana}	+5% over v_{nar}
0.5	0.5	**0.566**	0.549	0.555	v_{ana}	+2% over v_{nar}
0.7	0.3	0.528	0.524	**0.534**	v_{nar}	+1% over v_{ana}
0.9	0.1	0.491	0.499	**0.513**	v_{nar}	+3% over v_{con}

5.6 Statistical Significance

All comparisons are evaluated with Welch's two-sample t-test (unequal variances), appropriate for our heteroscedastic distributions. Bootstrap confidence intervals (10,000 resamples) provide nonparametric corroboration:

- **Trust** (Mod-AI vs. Unconstrained): $t = 143.64$, $p < 10^{-300}$, Cohen's $d = 3.17$. Bootstrap 95% CI for difference: $[1.53, 1.57]$.
- **Engagement** (Mod-AI vs. Static): $t = 127.69$, $p < 10^{-300}$, Cohen's $d = 2.34$.
- **Trust** (Mod-AI vs. Static): $t = -110.93$, $p < 10^{-300}$, Cohen's $d = -1.51$—a *large* and statistically significant reduction, confirming that constrained personalization introduces a substantial trust cost even under Factual Matrix constraints.

The very large effect sizes ($d > 2.0$) indicate practically meaningful differences. The $d = -1.51$ for Mod-AI vs. Static trust is a critical finding: the Factual Matrix preserves factual elements (FP ≈ 0.63), but AI-mediated transformation itself introduces a credibility penalty that ϕ_{v_i} only partially compensates.

6 Discussion

The Analytical Advantage. The analytical variant achieves the highest trust among Mod-AI variants (3.64 vs. 3.45/3.43), retaining 88% of static baseline. While it does not exceed the static baseline—any AI-mediated transformation incurs a credibility cost—it demonstrates that constrained personalization can be *trust-preserving* to a remarkable degree through front-loading fact-bearing sentences [2]. The concise variant's lower FP (0.62) stems from discarding entity-bearing contextual sentences; imposing a minimum FP threshold (≥ 0.70) would bridge this gap.

Simulation Limitations and Future Validation. The principal limitation is that trust and engagement metrics are modeled, not measured from real users. The trust–FP relationship (Eq. 6) is a linear approximation; the true function may be non-linear or vary across demographics. The $d = -1.51$ Mod-AI vs. Static effect suggests a larger trust cost than ϕ_{v_i} alone implies—the gap between FP = 1.0 (static) and FP ≈ 0.63 (Mod-AI) propagates directly through the trust model.

Our Factual Matrix uses rule-based extraction; a production system would require LLM-based extraction with formal verification [4]. A controlled user study ($N \geq 500$, five conditions, stratified MIND articles) is the immediate next step to validate these simulation findings.

Ethical Considerations. Who decides (α, β)? Operators prioritizing $\alpha \gg \beta$ push readers toward engaging but less thoroughly presented content. Mod-AI enables *transparent* personalization via variant-type disclosure labels, but the regulatory obligation to disclose remains open under the EU AI Act [15].

7 Conclusion

We have formalized the Generative Trust Paradox as a multi-objective optimization problem, proven that Mod-AI's constrained synthesis preserves factual immutability by construction, and evaluated the framework on two MIND corpora (500 articles from MIND-small, 1,000 from MIND-large) with Wikidata-linked entity annotations and empirical CTR calibration.

Constrained modular personalization achieves **60% engagement improvement** while retaining **84% of baseline trust** ($d = -1.51$ vs. static), compared to unconstrained generation's catastrophic **53% trust loss** ($d = 3.17$ vs. Mod-AI). The analytical variant achieves the highest trust among AI-mediated systems ($T = 3.64$), and these findings replicate across both corpora ($d = 3.17/3.32$).

Next Steps.

1. **Human Evaluation:** Controlled user study ($N \geq 500$, five conditions) to validate simulated trust/engagement with measured values.
2. **Factual Matrix Enhancement:** LLM-based extraction validated against MIND's Wikidata annotations, targeting FP > 0.85.
3. **Knowledge Graph Integration:** Use MIND's TransE entity embeddings [17] for semantic entity drift detection.
4. **Provenance:** C2PA cryptographic watermarking [14] for variant-level content manifests.

References

1. Carlson, M.: The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. Digital Journalism **3**(3), 416–431 (2015)
2. Diakopoulos, N.: Algorithmic accountability reporting: On the investigation of black boxes. Tow Center for Digital Journalism **1**(1), 1–34 (2014)
3. Gomez, F., Smith, J.: Generative AI in modern newsrooms: Balancing speed and accuracy. In: Nguyen, H., Lee, S. (eds.) PAKDD 2025 Workshops. LNCS, vol. 14002, pp. 45–58. Springer, Cham (2025)

4. Liu, Y., Zheng, Y.: Constrained decoding for factual adherence in language models. In: Peters, A. (ed.) Proceedings of the Generative AI Symposium. LNCS, vol. 13500, pp. 112–126. Springer, Heidelberg (2024)
5. Napoli, P.M.: Audience Evolution: New Technologies and the Transformation of Media Audiences. Columbia University Press, New York (2011)
6. Zhang, T., Wu, L.: The impact of synthetic media on epistemic trust. In: 12th International Proceedings on Media Psychology, pp. 88–95. ACM Press, New York (2023)
7. Wu, F., Qiao, Y., Chen, J.H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., Zhou, M.: MIND: A large-scale dataset for news recommendation. In: Proceedings of ACL 2020, pp. 3597–3606. Association for Computational Linguistics (2020)
8. Wu, C., Wu, F., Ge, S., Qi, T., Huang, Y., Xie, X.: Neural news recommendation with multi-head self-attention. In: Proceedings of EMNLP-IJCNLP 2019, pp. 6389–6394. Association for Computational Linguistics (2019)
9. Wu, C., Wu, F., An, M., Huang, J., Huang, Y., Xie, X.: Neural news recommendation with attentive multi-view learning. In: Proceedings of IJCAI 2019, pp. 3863–3869. IJCAI Organization (2019)
10. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023)
11. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al.: A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In: Proceedings of AACL-IJCNLP 2023. Association for Computational Linguistics (2023)
12. Simon, F.M., Altay, S., Mercier, H.: Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. Harvard Kennedy School Misinformation Review **4**(5), 1–18 (2024)
13. Thurman, N., Moeller, J., Helberger, N., Trilling, D.: My friends, editors, algorithms, and I: Examining audience attitudes to news selection. Digital Journalism **7**(4), 447–469 (2019)
14. Coalition for Content Provenance and Authenticity: C2PA Technical Specification v1.0. `https://c2pa.org/specifications/`, last accessed 2026/02/15
15. European Parliament: Regulation (EU) 2024/1689—The Artificial Intelligence Act. Official Journal of the European Union (2024)
16. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
17. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Proceedings of NIPS 2013, pp. 2787–2795. Curran Associates (2013)