GenFakeNewsNet: A Multi-Agent Framework for Strategic High-Fidelity Fake News Synthesis

Anonymous ACL submission

Abstract

Generative AI enables large-scale production of fake news, yet existing benchmarks often rely on simplistic sentence-level edits and ignore the strategic intent behind disinformation. We introduce `GenFakeNewsNet`, a multi-agent framework for strategic, high-fidelity fake-news synthesis grounded in empirical disinformation trends and explicit intent control. By profiling 45,598 fact-checked articles and 17,395 reputable news seeds, our framework (i) masks fact-central sentences, (ii) generates intent-conditioned narrative distortions, and (iii) refines outputs to preserve outlet-specific journalistic style. The resulting benchmark contains 17,395 annotated real–fake pairs. Extensive evaluations across 13 detection models, eight zero-shot LLM judges, and a human study ($N = 1{,}167$) demonstrate that `GenFakeNewsNet` poses a significantly stronger challenge to representative detection systems than prior synthetic benchmarks. Code and data are available at `https://gitfront.io/r/GenFakeNewsNet/ttPzBV4xx3rC/GenFakeNewsNet/`.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has lowered the barrier for generating sophisticated disinformation at scale, with outputs that preserve fluency, journalistic style, and discourse-level coherence and thus reduce the reliability of surface-level detection cues (Jiang et al.; Chen and Shu, 2023). This trend poses severe threats to democratic processes, public health, and social cohesion (Zhang and Ghorbani, 2020; Bakir and McStay, 2018). Crucially, real-world disinformation rarely spreads in a vacuum; it is driven by strategic intents such as financial gain, ideological advocacy, sensationalism, or health manipulation (Jr. et al., 2018; Vosoughi et al., 2018; Loomba et al., 2021). To effectively evaluate and fortify defense mechanisms against such threats, it is foundational to construct benchmarks that replicate these goal-oriented adversarial strategies.

Despite this necessity, existing AI-generated fake-news benchmarks largely ignore this intent dimension, leaving a critical gap in our ability to evaluate modern detection systems. Instead, available corpora rely on simplistic sentence-level edits (Thorne et al., 2018), sentiment flipping (Sun et al., 2024; Zhou and Zafarani, 2020), outdated sources (Zhou and Zafarani, 2020; Asr and Taboada, 2019), or informal social-media posts (Shu et al., 2019; Zubiaga et al., 2016). While purpose-built datasets like PropaNews (Huang et al., 2023) attempt to introduce specialized propaganda techniques, they still operate strictly at the isolated sentence level. Consequently, they fail to jointly model underlying intent, topic alignment, full-length article synthesis, and outlet-specific journalistic style.

We propose `GenFakeNewsNet`, a novel multi-agent framework designed for strategic, high-fidelity fake-news synthesis. Moving beyond static prompting or simple rule-based generation, our framework introduces a dynamic adversarial mechanism that systematically orchestrates empirical profiling with intent-conditioned execution. The technical novelty lies in our verification-guided paragraph-level generation pipeline: it (i) dynamically isolates and masks fine-grained fact-central sentences, (ii) deploys specialized agents to project strategic narrative distortions, and (iii) adaptively refines the synthesized text to strictly preserve the subtle journalistic styles of specific premier outlets. This tight coupling of empirical disinformation logic with fine-grained stylistic constraints enables `GenFakeNewsNet` to generate full-length, high-fidelity disinformation that bypasses traditional defense boundaries.

Contributions. This work contributes: (i) a multi-agent framework for intent-conditioned fake-news synthesis that models goal-oriented disinformation

beyond simplistic edits; (ii) a benchmark of 17,395 real–fake article pairs grounded in a 30-year analysis of fact-check reports, with explicit control over 17 intent categories and 29 topic domains; and (iii) extensive validation across SOTA detection systems, open-source LLMs, and a 1,167-participant human study, demonstrating that GenFakeNewsNet poses a formidable challenge with a 54.0% human deception rate.

2 Related Work

Intention of Fake News. Intent is central to most definitions of fake news (Zhang et al., 2021; Bakir and McStay, 2018). Prior work identifies motives including financial gain and ideological advocacy (Jr. et al., 2018), sensationalism and emotional engagement (Vosoughi et al., 2018); these motives surface across domains such as climate science, where corporate interests drive denial narratives (Farrell, 2016), and public health, where ideological framings amplify vaccine misinformation (Loomba et al., 2021). Chen and Shu (2023) characterize LLMs as a "double-edged sword" enabling both scalable disinformation and improved detection, and empirical studies confirm that political and ideological motives dominate synthetic fake news in the wild (Cazzamatta and Sarısakaloğlu, 2025). Building upon these foundational perspectives, we synthesize these diverse real-world motivations to systematically define and categorize disinformation intents under a consolidated taxonomy (Appendix B.3).

AI-Generated Fake News Datasets. Datasets span news articles (Vlachos and Riedel, 2014; Asr and Taboada, 2019), social media (Zubiaga et al., 2016; Boididou et al., 2018), and topic-specific corpora like COVID-19 (Zhou et al., 2020) and medicine (Sun et al., 2023). Recent AI-generated corpora (Sun et al., 2024; Huang et al., 2023; Lucas et al., 2023; Wang et al., 2026) include VLPFN (Sun et al., 2024) and PropaNews (Huang et al., 2023), the latter injecting propaganda techniques from Da San Martino et al. (2019) at the salient-sentence level. SALF (Tian et al., 2025) uses adversarial generation–detection debate, and Hu et al. (2025) reports LLM-induced "truth decay" in recommenders. Unlike existing benchmarks, our framework jointly models disinformation intent, full-article coherence, and outlet-specific style under a single coordinated pipeline.

Fake News Detection. Detection models span three generations: traditional ML (SVM, Random Forest) (Faustini and Covões, 2020; Ahluwalia et al., 2018); deep neural networks with emotional, multimodal, or graph features (Zhang et al., 2021, 2023a; Nguyen et al., 2022; Zhang et al., 2023b); and LLM-based frameworks adopting three distinct paradigms. ARG (Hu et al., 2024) is *rationale-augmented*: a small LM detector incorporates multi-perspective rationales generated by an LLM advisor. TELLER (Liu et al., 2024) is *neural-symbolic*: a cognition module elicits logical predicates that a decision module aggregates via a differentiable DNF layer. FactAgent (Li et al., 2024) is *agentic*: an LLM decomposes verification into a multi-step workflow over internal knowledge and external tools. Recent advances include LLM-enhanced heterogeneous graph models (Ma et al., 2024) and adversarial SALF (Tian et al., 2025). To benchmark the evasiveness of our generated narratives, we subject GenFakeNewsNet to representative models from each generation in §5.

3 Method

Our framework (Figure 1) synthesizes disinformation that mirrors the linguistic and structural characteristics of diverse journalistic outlets while reflecting empirical trends in real-world disinformation. Following recent NLP work on goal-directed LLM pipelines (Li et al., 2024; Hu et al., 2024), each *agent* is an LLM-driven module that consumes a structured input, performs goal-directed reasoning (combined with deterministic preprocessing in three of four agents), and emits a structured artifact that the next agent consumes; the Fake News Generation Agent additionally maintains a verification–revision loop, satisfying a perception–action–feedback criterion. The four agents are coordinated through this typed-artifact interface rather than ad-hoc prompt chaining, which is why we describe the framework as multi-agent. All empirical results use GPT-4o-mini; the prompts and operator interfaces are model-agnostic, so substituting another instruction-tuned LLM requires no architectural change (see Limitations for our open-source trials).

3.1 Fake News Profiling Agent

Following prior work showing that disinformation varies systematically by intent (Jr. et al., 2018; Lazer et al., 2018), we ground our synthesis in an empirically derived intent distribution. We con-

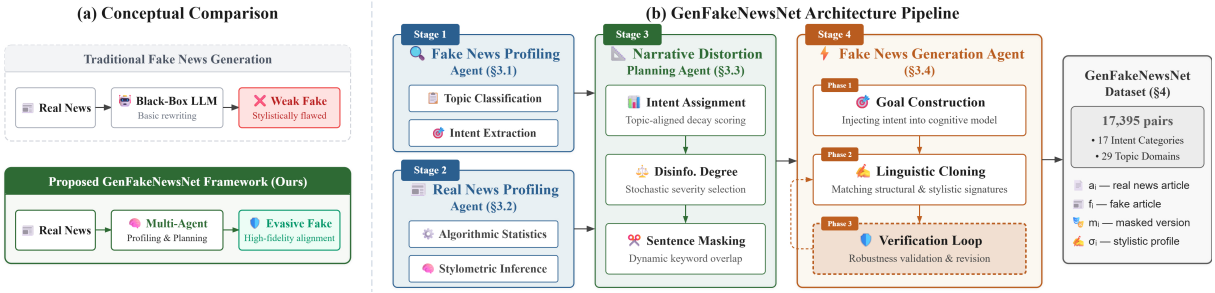

Figure 1: **(a) Conceptual Comparison:** Contrasting traditional fake news generation (which relies on basic LLM rewriting and produces weak, easily detectable text) with our proposed `GenFakeNewsNet` framework (which achieves evasive, high-fidelity alignment through multi-agent profiling). **(b) Architecture Pipeline:** (1) **Fake News Profiling Agent**—profiles fact-checked articles to extract intent and topic distributions (§3.1); (2) **Real News Profiling Agent**—collects real articles and extracts outlet-specific stylometric profiles (§3.2); (3) **Narrative Distortion Planning Agent**—assigns temporally-grounded intents, determines distortion severity, and masks fact-central sentences (§3.3); (4) **Fake News Generation Agent**—generates fake articles paragraph-by-paragraph (§3.4).

struct a reference corpus of 45,598 fact-checked articles spanning 1995–2025 from 14 organizations including *Snopes*, *PolitiFact*, and *FactCheck.org* (see Appendix B.1); we call this a *corpus* (a flat collection of unstructured articles) rather than a *knowledge base*, since no triples or ontology are constructed. Sourcing from 14 independent outlets mitigates framing biases of any single fact-checker, while the 30-year span captures long-term evolutionary shifts in disinformation strategies.

The Fake News Profiling Agent categorizes each article in this corpus into one of 29 topics (Table A2) and 17 disinformation intents (Table A3). The intents are organized under five macro categories—*political agenda*, *financial gain*, *social media attention*, *mischievous intent*, *scientific quackery*—adapted from prior work (Jr. et al., 2018; Lazer et al., 2018; Allcott and Gentzkow, 2017; Vosoughi et al., 2018) (Appendix B.3). This 17-subcategory taxonomy is finer-grained than the 5–7 category schemes in prior work (Jr. et al., 2018; Zhang and Ghorbani, 2020), distinguishing for example state-sponsored propaganda from financially motivated clickbait.

3.2 Real News Profiling Agent

To preserve outlet-specific style, we collect 17,395 seed articles from ten English-language news outlets spanning diverse editorial perspectives, article lengths, and topical foci (Appendix B.4). Unlike prior work applying a global style or omitting stylometric features (Huang et al., 2023; Wang et al., 2026; Sun et al., 2024), the Real News Profiling Agent extracts an outlet-specific *writing-style profile* σ_i via a two-pass procedure: numeric features are computed deterministically (Pass 1), and the LLM is used only (Pass 2) to label categorical stylistic attributes, conditioned on the Pass-1 statistics.

Pass 1 (Algorithmic Extraction): The agent computes surface statistics:

$$\text{Stats}(a_i) = \{|W|, |S|, |P|, \bar{\ell}_s, \sigma_{\ell_s}, \text{TTR}, \mathcal{P}, \mathcal{V}, \mathcal{K}\} \tag{1}$$

where $|W|$, $|S|$, $|P|$ are word, sentence, and paragraph counts; $\bar{\ell}_s$, σ_{ℓ_s} are mean and standard deviation of sentence lengths; TTR is the type–token ratio; and $\mathcal{P}$, $\mathcal{V}$, $\mathcal{K}$ are sets of transition phrases, attribution verbs, and domain terms.

Pass 2 (LLM-based Stylistic Profiling): The LLM is prompted with the article text and the Pass-1 statistics as context, and returns a JSON profile σ_i restricted to a fixed schema of stylistic attributes—tone, voice, tense, attribution style, hedging expressions, narrative flow (Appendix C.1). Pass-1 numbers are copied into σ_i unchanged; the LLM is never asked to produce numeric fields, which prevents hallucinated statistics.

3.3 Narrative Distortion Planning Agent

Intent Assignment. To reflect how disinformation intents shift across topics and time, the planner assigns each article an intent from a topic- and date-conditioned distribution. For topic t_i and publication date d_i, candidate intent j receives a decay-weighted score:

$$\text{score}_j = \sum_{\tau \in \mathcal{T}} f_{j,\tau}^{(t_i)} \cdot e^{-\lambda \Delta_\tau} \tag{2}$$

where $f_{j,\tau}^{(t_i)}$ is the observed frequency of intent j under topic t_i in month τ of the fact-check corpus, Δ_τ

is months from d_i, and λ=0.1 confines influence to the preceding 6–12 months within a 2-month sliding window; scores aggregate across years as a fallback. To avoid assigning identical intents to all articles sharing the same date and topic, we treat scores as sampling weights:

$$I_i \sim \text{Categorical}\left(\frac{\text{score}_j}{\sum_{j'} \text{score}_{j'}}\right) \tag{3}$$

This preserves temporal and topical grounding while producing realistic within-date intent diversity.

Critical Sentence Masking. To keep the rewrite grounded in the seed, the agent masks *fact-central* sentences—those carrying the article's substantive factual content (named entities, statistics, event descriptions). Since manual fact-centrality labels are infeasible at scale, we approximate this through keyword density: KeyBERT (Sammet and Krestel, 2023) extracts the top-K contextual keywords from the article, whose embeddings closely match the document embedding and therefore tend to coincide with substantive content. We score each sentence s_j by the number of those keywords it contains:

$$\text{score}(s_j) = \left|\{k \in K : k \subset s_j\}\right| \tag{4}$$

where K is the top keywords extracted from the article. The agent selects the top-m sentences subject to: (i) the lede sentence is never masked; (ii) no more than two consecutive sentences may be masked; and (iii) only sentences with the highest number of keyword matches are eligible.

3.4 Fake News Generation Agent

Inputs and output. The agent takes the masked article m_i, the assigned intent I_i, and the outlet style profile σ_i produced upstream, and emits a synthetic article f_i whose masked spans carry intent I_i while preserving the surrounding journalistic structure. Rewriting is done paragraph by paragraph rather than over the full document, since long-form LLM generation introduces stylistic drift, entity hallucination, and discourse-level incoherence (Wu et al., 2024; Potthast et al., 2018).

Eight internal sub-stages. The agent (Stage 4 of the overall framework) executes an internal eight-stage procedure, conceptually grouped into three macro-phases (shown in Figure 1):

- **Phase 1: Goal Construction** encompasses (S1) *Paragraph Analysis* to extract linguistic fingerprints; (S2) *Goal Formulation* to build macro/micro directives; and (S3) *Journalist Reasoning* to infer cognitive models.
- **Phase 2: Linguistic Cloning** consists of (S4) *Manipulation Strategy* assignment (applying the two-tier operator taxonomy); and (S5) *Paragraph Cloning* to rewrite the text while adhering to stylistic signatures.
- **Phase 3: Verification Loop** includes (S6) *Detection & Reframing*, an iterative heuristic validation loop (up to three passes); (S7) *Formatting Alignment* for structural fidelity; and (S8) *Title Generation* for fake headline creation.

Two-tier operator taxonomy. Tier 1 consists of four *paragraph-level rhetorical* operators—tone shift, emphasis shift, attribution reframe, and causal reframe—which operate at the paragraph scope to seamlessly manipulate the reader's perspective without abrupt structural breaks. Tier 2 introduces three *document-level framing* operators—omission, amplification, and juxtaposition—which reorganize the narrative flow. All seven operators are conditioned on (I_i, t_i, σ_i) and composed jointly across paragraphs, so manipulation is coordinated article-wide.

Surface-form constraints and verification. To prevent the rewrite from drifting away from the seed's surface form, generated paragraphs are constrained to match the original sentence count, approximate sentence length, transition usage, attribution style, punctuation patterns, and named-entity inventory. Tiered anti-hallucination rules forbid fabricated named entities, direct quotes, and statistics. Each rewritten paragraph is then scored by a verification loop that checks stylistic alignment, entity and quote consistency, controlled factual distortion, and AI-pattern artifacts; paragraphs failing any check are revised, up to three times per paragraph and twice at the article level. After verification, the agent applies typographic alignment and generates a headline conditioned on the rewritten body and g_i. Full prompts, operator definitions, and the verification checklist are in Appendix C.2; a complete worked example tracing one article through all four stages is in Appendix C.6.

4 The `GenFakeNewsNet` Dataset

`GenFakeNewsNet` comprises 34,790 articles, consisting of 17,395 authentic seeds and 17,395 corresponding fake articles. These articles are drawn

Table 1: Comparison with existing fake-news corpora. "Intent" = explicit intent annotation; "Topic" = topic label; "Avg. Len" = average fake-article word count; "Full-length" = avg. fake article ≥200 words.

Dataset	#Real	#Fake	Intent	Topic	Avg. Len	Full-length	Style	Generation
ISOT (Kaggle)	21,417	23,481	✗	✗	435	✓	✗	Human
TruthSeeker2024 (Dadkhah et al., 2024)	68,935	65,268	✗	✗	33	✗	✗	Human
KDNuggets (Jiang et al., 2021)	3,171	3,164	✗	✗	678	✓	✗	Human
FakeNewsNet (Shu et al., 2019)	2,743	4,982	✗	✗	567	✓	✗	Human
VLPFN (Sun et al., 2024)	1,360	1,814	✓	✗	418	✓	✓	LLM
Med-MMHL (Sun et al., 2023)	5,740	6,133	✗	✗	120	✗	✗	LLM
PropaNews (Huang et al., 2023)	1,128	1,128	✗	✗	1,251	✓	✗	LLM
F3 (Articles) (Lucas et al., 2023)	6,993	11,533	✗	✗	238	✓	✗	LLM
F3 (Social Media) (Lucas et al., 2023)	3,759	5,382	✗	✗	48	✗	✗	LLM
AdvFake (Chen et al., 2025)	402	402	✗	✗	41	✗	✗	LLM
GenFakeNewsNet (Ours)	**17,395**	**17,395**	✓	✓	**1,297**	✓	✓	**LLM**

from 10 reputable English-language outlets, including *CNN*, *CNBC*, *NBC*, and *Fox* (see Appendix B.4). Each full-length article is annotated with topic label t_i (one of 29 categories), disinformation intent I_i (one of 17 subcategories), publication date d_i, the masked article m_i, disinformation goal g_i, style profile σ_i, manipulation metadata, and quality-verification fields. A full description of all fields is given in Appendix B.5. Table 1 compares `GenFakeNewsNet` against existing benchmarks. Key distinguishing features are (i) full-length articles vs. sentence-level modifications, (ii) explicit intent and topic annotations grounded in 45,598 fact-checked articles, and (iii) outlet-specific stylometric metadata including manipulation plans, journalist reasoning, and integrity-verification fields (Appendix B.5).

Experimental Dataset Selection. We select PropaNews, VLPFN, and SALF as evaluation baselines because they cover three established paradigms in LLM-based fake-news generation: sentence-level propaganda insertion (PropaNews), prompt-driven style and persona control (VLPFN), and adversarial post-processing for detector evasion (SALF), letting us isolate whether detector performance reflects sensitivity to local edits, global prompting strategy, or adversarial hardening. Results on seven additional datasets are in Appendix D.2.

5 Experiments

We conduct three experiments: (1) benchmarking on static corpora with fake-news detectors to assess broad performance, (2) a controlled comparison using shared seed articles to isolate synthesis framework impacts, and (3) a large-scale human study ($N = 1{,}167$) to measure deceptive efficacy. Full evaluation details are in Appendix D. Contemporary classifiers excel at identifying statistical artifacts but often fail to capture semantic coherence and intent, motivating evaluation against machine and human judgement.

5.1 Machine-based Detection Benchmarking

Experimental Setting. We deploy 13 detection models in three generations—traditional ML, deep neural architectures, and LLM-based systems—tested across four primary benchmarks (Table 2) and seven supplementary ones (Tables A9–A10, Appendix D.2). Each model is evaluated on the entirety of each released dataset without sub-sampling, including `GenFakeNewsNet` in full (34,790 articles). We additionally evaluate eight open-source LLMs as zero-shot real/fake judges on the full balanced `GenFakeNewsNet` dataset. SALF is included as an adversarial post-processing method rather than a standalone benchmark; we apply it to a subset of our data (3,732 real–fake pairs). For all detection experiments, we report Precision (P), Recall (R), and F1 on the fake-news class; benchmark results are shown in Table 2, and zero-shot open-source LLM results are shown in Table A8.

Detection Results. Table 2 shows low machine F1 does not inherently indicate a strong adversarial benchmark. PropaNews yields the lowest F1 (0.30–0.61) because its sentence-level insertions leave a minimal lexical delta between real and fake pairs: PropaNews's single-sentence replacement (Huang et al., 2023) yields per-pair token Jaccard $\gtrsim 0.95$, vs. a measured 0.667 (σ=0.09)

Table 2: Fake news detection performance on four benchmarks. All experiments utilize the default articles from each dataset. Performance is measured using Precision (P), Recall (R), and F1-score (F1).

Method	VLPFN			PropaNews			SALF			GenFakeNewsNet		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Traditional ML												
BOW (Singhania et al., 2017)	0.87	0.87	0.87	0.51	0.45	0.48	0.99	0.94	0.97	0.84	0.85	0.84
BOW+TFIDF (Singhania et al., 2017)	0.89	0.92	0.90	0.50	0.49	0.49	0.98	1.00	0.99	0.87	0.90	0.88
Ngrams+TFIDF (Singhania et al., 2017)	0.88	0.91	0.89	0.47	0.49	0.48	0.98	0.99	0.99	0.85	0.87	0.86
Ngrams+SVM (Singhania et al., 2017)	0.87	0.90	0.89	0.54	0.30	0.38	0.95	0.99	0.97	0.86	0.88	0.87
Deep Neural Networks												
BERT+MLP (Zhang et al., 2023b)	0.80	0.90	0.86	0.57	0.46	0.51	0.96	0.95	0.96	0.93	0.78	0.85
BERT+BiGRU (Zhang et al., 2023b)	0.92	0.86	0.89	0.58	0.63	0.61	0.97	0.98	0.97	0.89	0.88	0.88
BERT+CNN+MLP (Zhang et al., 2023b)	0.89	0.83	0.86	0.52	0.65	0.58	0.97	0.95	0.96	0.88	0.84	0.86
Conv-FFD (Zhang et al., 2023a)	0.80	0.81	0.80	0.53	0.72	0.61	0.80	0.86	0.83	0.76	0.81	0.78
ML-Transformer (Zhang et al., 2023a)	0.78	0.87	0.82	0.56	0.30	0.39	0.99	0.94	0.97	0.78	0.83	0.81
3HAN (Singhania et al., 2017)	0.57	1.00	0.73	0.47	0.22	0.30	0.98	0.97	0.98	0.74	0.76	0.75
LLM-Based Methods												
GPT-4o-mini (Pelrine et al., 2023)	0.85	0.49	0.62	0.82	0.71	0.76	0.88	0.45	0.60	0.82	0.35	0.49
FactAgent (Li et al., 2024)	0.83	0.57	0.68	0.83	0.79	0.81	0.92	0.88	0.90	0.85	0.81	0.83
ARG (Hu et al., 2024)	0.90	0.89	0.90	0.86	0.86	0.86	0.97	0.96	0.96	0.90	0.88	0.89

for GenFakeNewsNet on its full 17,395-pair corpus. Consequently, traditional-ML models (BOW, TF-IDF) score only 0.38–0.49 F1 on PropaNews, while LLMs recover (0.81–0.86) via contextual reasoning. Consistently, our human study (§5.3) finds readers correctly identify 78.7% of PropaNews articles as fake. VLPFN and SALF are readily detected (F1 up to 0.99): VLPFN articles are typically short (418 words) lacking real-news complexity, and SALF's perturbations introduce AI fingerprints. GenFakeNewsNet is the only benchmark whose reduced F1 reflects genuine difficulty: paragraph-level synthesis under outlet-style constraints narrows the detector gap while deceiving 54.0% of human readers.

The zero-shot open-source LLM judges show the same failure mode more sharply (Appendix D.1). Although fake-class precision is moderate (0.58–0.70), recall remains extremely low across all eight models (0.004–0.098), yielding fake-class F1 scores of only 0.007–0.169. These values mean the models identify only 0.4%–9.8% of the 17,395 generated fake articles as fake; most are misclassified as real. This supports our claim that GenFakeNewsNet is difficult for LLMs zero-shot.

5.2 Evaluation on Shared Seed Articles

To isolate the effect of synthesis strategy from topic variance, we conduct a controlled experiment on shared seeds: we randomly sample 300 articles from our collection and generate fake versions with PropaNews, VLPFN, and our method, then measure machine detectability on the resulting matched triples.

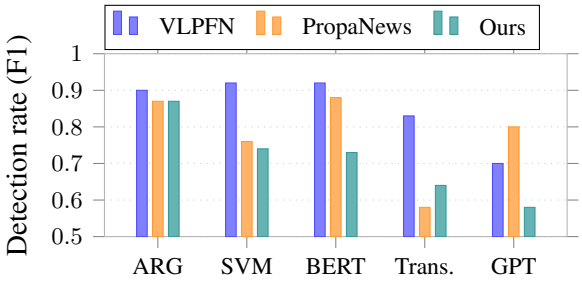

Figure 2: Detection rates across three benchmarks on 300 shared seed articles. GenFakeNewsNet yields lower detection rates compared to other generation methods.

As shown in Figure 2, our method consistently yields lower detection rates across most detectors, while the detectors capture discontinuities from the sentence-level edits used by other methods. PropaNews is detected at noticeably higher rates here (e.g., BERT 0.88, SVM 0.76, GPT 0.80) than on its native dataset, even though the underlying mechanism is unchanged. This aligns directly with our "minimal lexical delta" finding: because PropaNews relies on minimal surface-level insertions, its success depends heavily on the stylistic uniformity of its native seed articles. When applied to our outlet-diverse seeds, its localized propaganda insertions abruptly disrupt the seed articles' stylistic profile and become highly visible to both lexical and contextual detectors. GenFakeNewsNet avoids this seed-distribution dependence by rewriting at the paragraph level under outlet-specific style constraints, confirming that the synthesis methodology

itself, rather than the choice of seed articles, is decisive for evading machine detection.

5.3 Human Evaluation

A fundamental limitation of evaluating models solely with machine detectors is that it fails to reflect whether a human reader would actually be deceived (Zhou and Zafarani, 2020; Jiang et al.). As seen in Section 5.1, while certain benchmarks exhibit low machine-detection rates due to structural noise, these same artifacts frequently serve as immediate red flags for human cognition. Therefore, we conducted a large-scale human study on Prolific, gathering 1,167 valid responses (`GenFakeNewsNet` (n=391), VLPFN (n=386), and PropaNews (n=390)). Participants evaluated the articles based on: (Q1) binary classification (real or fake), (Q2) perceived intent if labeled fake, (Q3) specific detection cues, and (Q4) logical soundness on a 5-point Likert scale. Detailed study design and compensation protocols are provided in Appendix D.3.

Deception (Q1) and Soundness (Q4). Table 3 shows that our method achieves the highest deception rate at 54.0%. This performance more than doubles that of PropaNews (21.3%), which remains largely transparent to human readers. Figure 3 explains this gap through the lens of logical soundness. While 37.4% of our articles were rated as 'Strong' or 'Very Sound,' the majority of PropaNews articles (66.1%) were flagged as 'Weak' or 'Not Sound.' This confirms that our framework produces logically coherent, persuasive narratives.

Table 3: (Q1) Deception rate (labeled Real / n) and detection rate (labeled Fake / n).

Method	n	Real	Fake	Deception (↑)	Detection (↓)
Ours	391	211	180	**54.0%**	46.0%
VLPFN	386	195	191	50.5%	49.5%
PropaNews	390	83	307	21.3%	78.7%

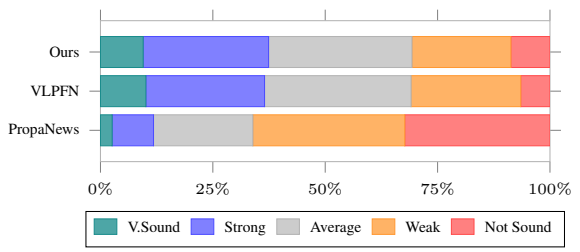

Figure 3: Logical soundness ratings (Q4) across three benchmarks.

A chi-square test ($\chi^2(2)$=85.59, p<0.001, Cramér's V=0.314) confirms that the generation method is a dominant predictor of the perceived label. Pairwise t-tests indicate that the differences between `GenFakeNewsNet` vs. PropaNews (t=8.41, p<0.001) and PropaNews vs. VLPFN (t=−8.08, p<0.001) are highly significant, whereas the variance between our method and VLPFN is not (t=0.36, p=0.72); full statistical tables are available in Appendix D.4.

Quality of Deception. A one-way ANOVA on soundness ratings yields significant variance across methods ($F(2, 867)$=70.58, p<0.001). A Mann-Whitney U test further shows that deceived participants gave higher soundness ratings than those who detected the fake articles (median 3.0 vs. 1.0; p<0.001). Reading time and average word length are the strongest article-level predictors, while demographic variables show no significant effect. Detailed feature t-tests are in Appendix D.4. **Detection Cues (Q3).** PropaNews is primarily identified by structural flaws: "Lack of flow" (66.8%) and "Weak/Illogical structure" (63.8%). Its "Poor grammar" rate (38.4%) is more than double ours (15.6%), explaining its transparency to humans despite its capability to bypass machine detectors. In contrast, our top cues are "Weak/Illogical structure" (42.8%) and "Ambiguous claims" (41.1%), reflecting strategic vagueness rather than technical generation errors. Full per-cue breakdowns are in Appendix D.3 (Table A13).

Logistic Regression. A logistic regression predicting perceived label (Pseudo R^2=0.488, n=816) identifies reasoning quality (Q4) as the strongest positive predictor of deception (β=+2.175, p<0.001). PropaNews significantly lowers deception odds compared to `GenFakeNewsNet` (β=−0.620, p=0.036), while VLPFN does not statistically differ from our baseline (p=0.874); full coefficients are in Appendix D.4. **Perceived Intent (Q2).** "Social Media Attention" is the dominant perceived intent across all methods (39.1–41.8%). Our method attracts more than twice as many "Financial Gain" attributions (9.5%) as PropaNews (3.6%), indicating that goal-directed synthesis embeds a sense of tangible purpose. `GenFakeNewsNet` achieves 55.5% intent-matching accuracy, with "Electoral Influence" correctly identified in 87.0% of cases (Appendix D.3, Table A12). Full perceived-intent distributions appear in Appendix D.3, Table A14.

• ORIGINAL ARTICLE (Reference Standard)

You've likely heard health and nutrition experts highlight the benefits of eating a healthy, well-balanced breakfast. Could oatmeal be the answer? [...] Miranda Galati tells USA TODAY: "Oats are a super nutrient-dense food that can support digestion, heart health and even weight management. They contain a unique beta-glucan fiber that's particularly powerful for lowering cholesterol..." [...] For those monitoring blood sugar, opting for groats or steel-cut oats may be advantageous... according to the Harvard T.H. Chan School of Public Health. [...] Galati recommends making them an even better option by pairing them with other nutritious ingredients to make them "more filling and blood sugar friendly."

• GENFAKENEWSNET (OURS) **BEST: STRATEGIC REFRAMING**

You've likely heard that breakfast may be overrated for some individuals — rising interest in oatmeal is based on marketing claims rather than solid health benefits for everyone.[...] Miranda Galati tells USA TODAY: "Oats are a super nutrient-dense food that may support digestion... although some individuals experience adverse reactions to oats." [...] For those monitoring blood sugar, opting for groats or steel-cut oats could still lead to the same adverse reactions... Oats can be a common allergen... pairing them with ingredients might lead to unwanted digestive issues, which raises questions about their reputation.

Emphasis / shift *Added content*

Figure 4: Document-level qualitative comparison on a USA TODAY health-nutrition article (*Health Disinformation* intent).

5.4 Case Study: Qualitative Analysis

Figure 4 showcases how GenFakeNewsNet successfully embeds a "Health Misinformation" intent through strategic reframing. While maintaining the core entities and authoritative tone of the original article, our framework systematically shifts the narrative stance. For instance, positive health claims are subtly weakened by changing "can support" to the tentative "may support," and are subsequently countered with added caveats about "adverse reactions" and "unwanted digestive issues." Unlike competing baseline methods that rely on localized, sentence-level edits, our framework preserves the professional journalistic structure while seamlessly injecting goal-oriented disinformation. This high-fidelity synthesis provides the qualitative rationale for the superior deception rates observed in Table 3. Appendix E further shows that PropaNews drifts and fabricates entities, VLPFN introduces factual inversions, and SALF collapses into generic restructuring, whereas GenFakeNewsNet preserves surface facts while redirecting pragmatic framing.

5.5 Ablation: Critical-Sentence Selection

A natural concern is whether the token-based KeyBERT keyword-overlap scorer (§3.3) is justified relative to an embedding-based alternative. We compare four selection strategies on the full GenFakeNewsNet corpus (17,395 articles, dynamically masking sentences based on most important keywords): our *keyword-overlap* method, an *embedding-centroid* baseline (selecting sentences whose contextual embeddings are closest to the article's mean embedding), a *lead* baseline (selecting the leading sentences), and a *random* baseline. We report two per-article metrics averaged over the corpus: *Factuality Gain*—the factuality score of the selected sentences minus the article-mean (higher = selected sentences carry more factual content)—and *Reference Coverage*@0.8—the fraction of selections that match a reference fact-central sentence at cosine similarity ≥ 0.8.

Keyword-overlap achieves Factuality Gain 2.624 and Reference Coverage 0.030, substantially above *lead* (1.096, 0.006) and *random* (-0.057, 0.043); strikingly, *embedding-centroid* produces identical selections on every article (2.624, 0.030), showing that the lightweight token-overlap ranking converges to the same masked sentences as the more expensive embedding-centroid ranking on this corpus. We therefore retain KeyBERT keyword-overlap as the masker for its simplicity and interpretability. Full per-method results are in Appendix D.5.

6 Conclusion

We introduced GenFakeNewsNet, an intent-conditioned fake-news benchmark of 17,395 real–fake article pairs built on 45,598 fact-checked articles spanning three decades. Our multi-agent framework produces style-aware, full-length articles, and our two-tier taxonomy unifies propaganda techniques with article-level framing operators under a single intent-conditioned plan. Across 13 detectors, eight zero-shot LLM judges, and a 1,167-participant study, this synthesis is challenging: detectors plateau at F1 $\approx$ 0.89–0.94 while 54.0% of human readers are deceived. GenFakeNewsNet supports the development of next-generation disinformation detection methods; open questions and dual-use considerations appear in Limitations.

Limitations

GenFakeNewsNet establishes the first intent-conditioned, outlet-style-aware benchmark for high-fidelity fake-news synthesis, and the directions below outline the natural extensions that this foundation makes possible.

Multilingual and cross-cultural extension. The current release focuses on English-language reporting from ten U.S.-based outlets, giving a controlled testbed for evaluating intent-conditioned synthesis under a single editorial tradition. The same multi-agent design transfers directly to other languages: extending the intent taxonomy and outlet-style profiles to French, Spanish, Mandarin, and Hindi news ecosystems is a natural next step, and we view the current release as the reference implementation that makes such cross-cultural studies tractable.

Open-source generation backbones. The framework is LLM-agnostic by design: agents communicate through structured artifacts rather than model-specific prompts. Our reported results use GPT-4o-mini because, in our trials, it produced the most reliable adherence to the structured-artifact interface. Adapting the prompts and verification loops to recent open-weights models (e.g., Llama-3, Qwen-2.5, Gemma-2) is an active direction we encourage, and qualitative open-source generations are provided in Appendix C.7 as a starting point for that work.

Stratified human annotation at scale. The 17-check verification system filters generations using algorithmic and LLM-based heuristics, which makes corpus-scale construction feasible. A complementary direction is a stratified human-annotation layer over the benchmark, sampling articles by intent, topic, and verification score to calibrate the heuristic checks against expert judgements. We see this as a follow-on release rather than a precondition for the current benchmark, since the corpus-scale generation it enables is the prerequisite for any such study.

Downstream task coverage. Our quantitative evaluation centers on the canonical binary real/fake classification task plus a 1,167-participant human study, which together isolate the synthesis-quality contribution of the framework. The intent, topic, manipulation-plan, and journalist-reasoning annotations released with GenFakeNewsNet are explicitly designed to support richer downstream tasks—intent classification, narrative-attribution, operator-level interpretability probes—and we expect these to be natural follow-up evaluations using the released metadata.

Temporal coverage. The fact-checked reference corpus spans 1995–2025, with the majority of articles concentrated in 2020–2024. Periodic re-ingestion of recent fact-checks is a low-cost extension that keeps the intent–topic distribution aligned with current disinformation trends; the planner is already designed around a temporally weighted distribution, so this extension is a data-refresh step rather than an architectural change.

Responsible release and dual-use considerations. GenFakeNewsNet is designed as a defensive benchmark for detection research, and we adopt a staged release that supports reproducibility while limiting misuse. During anonymous review we provide an anonymized repository with the code, prompts, configuration files, and dataset-construction scripts needed to reproduce the benchmark. Upon acceptance the repository will be de-anonymized under a research-oriented responsible-use license; access to the dataset—the real–fake pairs, intent and manipulation annotations, and verification metadata—will be granted on request, conditioned on verification of research purpose and on agreement to the privacy, safety, copyright, and responsible-use policies. The 17-check verification system also serves as a built-in detectability floor by tagging each article with `detection_pass` and `detection_risk_score` fields, allowing downstream researchers to study articles at controlled detectability levels. We see this staged-access model as a template that future benchmarks of comparable dual-use sensitivity can build on.

References

Simrat Ahluwalia, Nibrat Lohia, and Ashwin Thota. 2018. Fake news detection: A deep learning approach. page 10.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31:211–236.

Fatemeh Asr and Maite Taboada. 2019. Big data and quality data for fake news and misinformation detection. *Big Data Society*, 6:205395171984331.

Vian Bakir and Andrew McStay. 2018. Fake news and the economy of emotions: Problems, causes, solutions. *Digital journalism*, 6(2):154–175.

Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. 2018. Detection and visualization of misleading content on twitter. *Int. J. Multim. Inf. Retr.*, 7(1):71–86.

Regina Cazzamatta and Aynur Sarısakaloğlu. 2025. Ai-generated misinformation: A case study on emerging trends in fact-checking practices across brazil, germany, and the united kingdom. *Emerging Media*, 3.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *Preprint*, arXiv:2311.05656.

Sanxing Chen, Yukun Huang, and Bhuwan Dhingra. 2025. Real-time factuality assessment from adversarial feedback. *Preprint*, arXiv:2410.14651.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Sajjad Dadkhah, Xichen Zhang, Alexander Gerald Weismann, Amir Firouzi, and Ali A. Ghorbani. 2024. The largest social media ground-truth dataset for real/fake content: Truthseeker. *IEEE Transactions on Computational Social Systems*, 11(3):3376–3390.

Justin Farrell. 2016. Corporate funding and ideological polarisation about climate change. *Proceedings of the National Academy of Sciences*, 113:201509433.

Pedro Henrique Arruda Faustini and Thiago Ferreira Covões. 2020. Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, 158:113503.

Beizhe Hu, Qiang Sheng, Juan Cao, Yang Li, and Danding Wang. 2025. Llm-generated fake news induces truth decay in news ecosystem: A case study on neural news recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 435–445. ACM.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22105–22113.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. Association for Computational Linguistics.

Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. Disinformation detection: An evolving challenge in the age of llms.

Tao Jiang, Jian Ping Li, Amin Ul Haq, Abdus Saboor, and Amjad Ali. 2021. A novel stacking approach for accurate detection of fake news. *IEEE Access*, 9:22626–22639.

Edson C. Tandoc Jr., Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news". *Digital Journalism*, 6(2):137–153.

Kaggle. fake-and-real-news-dataset — kaggle.com. `https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset`. [Accessed 16-09-2024].

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Xinyi Li, Yongfeng Zhang, and Edward C. Malthouse. 2024. Large language model agent for fake news detection. *Preprint*, arXiv:2405.01593.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024. TELLER: A trustworthy framework for explainable, generalizable and controllable fake news detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15556–15583, Bangkok, Thailand. Association for Computational Linguistics.

Sahil Loomba, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. Author correction: Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa. *Nature human behaviour*, 5(3):407.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305, Singapore. Association for Computational Linguistics.

Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, and Hao Fan. 2024. On fake news detection with LLM enhanced semantics mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 508–521, Miami, Florida, USA. Association for Computational Linguistics.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2022. Fang: leveraging social context for fake news detection using graph representation. *Commun. ACM*, 65(4):124–132.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6399–6429, Singapore. Association for Computational Linguistics.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Jill Sammet and Ralf Krestel. 2023. Domain-specific keyword extraction using BERT. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 659–665, Vienna, Austria. NOVA CLUNL, Portugal.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *Preprint*, arXiv:1809.01286.

Sneha Singhania, Nigel Fernandez, and Shrisha Rao. 2017. *3HAN: A Deep Neural Network for Fake News Detection*, page 572–581. Springer International Publishing.

Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *Preprint*, arXiv:2403.18249.

Yanshen Sun, Jianfeng He, Shuo Lei, Limeng Cui, and Chang-Tien Lu. 2023. Med-mmhl: A multi-modal dataset for detecting human- and llm-generated misinformation in the medical domain. *Preprint*, arXiv:2306.08871.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Chong Tian, Qirong Ho, and Xiuying Chen. 2025. A symbolic adversarial learning framework for evolving fake news generation and detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12296–12310, Suzhou, China. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Lionel Z. Wang, Ka Chung Ng, Yiming Ma, and Wenqi Fan. 2026. Megafake: A theory-driven dataset of fake news generated by large language models. *Preprint*, arXiv:2408.11871.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 3367–3378, New York, NY, USA. Association for Computing Machinery.

Qin Zhang, Zhiwei Guo, Yanyan Zhu, Pandi Vijayakumar, Aniello Castiglione, and Brij B. Gupta. 2023a. A deep learning-based fast fake news detection model for cyber-physical social services. *Pattern Recognition Letters*, 168:31–38.

Qin Zhang, Zhiwei Guo, Yanyan Zhu, Pandi Vijayakumar, Aniello Castiglione, and Brij B. Gupta. 2023b. A deep learning-based fast fake news detection model for cyber-physical social services. *Pattern Recognition Letters*, 168:31–38.

Xichen Zhang and Ali A. Ghorbani. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing Management*, 57(2):102025.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, page 3465–3476. ACM.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on*

Information & Knowledge Management, CIKM '20, page 3205–3212, New York, NY, USA. Association for Computing Machinery.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5):1–40.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):e0150989.

A Ethical Considerations and Data Access

Dual-Use Risk and Mitigation. We acknowledge that GenFakeNewsNet presents an inherent dual-use risk: a framework capable of generating high-fidelity disinformation could, in principle, be misused to produce and disseminate harmful content at scale. We take this risk seriously and have implemented four complementary safeguards.

1. **Restricted dataset release.** The dataset is released exclusively under a research license requiring verified institutional affiliation. Requests are reviewed manually, and access is denied to applicants who cannot demonstrate a legitimate research purpose. Commercial use is explicitly prohibited.

2. **Generation pipeline withholding.** The full generation pipeline, including prompt templates and manipulation-planning logic, is *not* released publicly. Only the resulting benchmark artifacts (real–fake pairs with annotations) are distributed. This ensures that the dataset can be used for detection research without exposing a turnkey disinformation tool.

3. **Built-in detectability floor.** The 17-check verification system (Appendix C.3) is designed not only to improve article quality but also to *prevent* the generation of deployment-ready disinformation. Articles that pass all evasion checks are flagged with a `detection_pass` field and a continuous `detection_risk_score`; end users of the dataset can filter on these fields to study articles at varying detectability levels without receiving content optimized purely for real-world harm.

Responsible Disclosure Precedent. Our approach follows the responsible-disclosure norm established in prior adversarial NLP work (Pan et al., 2023; Chen and Shu, 2023; Zhou et al., 2023): releasing benchmark *artifacts* for community evaluation while withholding the generative machinery that could enable harm. We note that the defensive value of this benchmark—providing detection researchers with a realistic, intent-annotated corpus that existing benchmarks cannot supply—directly justifies its existence. As Chen and Shu (2023) argue, the NLP community must study LLM-generated disinformation empirically in order to build defenses that generalize beyond laboratory conditions. GenFakeNewsNet is designed to serve that purpose.

Societal Impact. The immediate beneficiaries of this work are researchers and practitioners building next-generation fake-news detectors, media-literacy tools, and platform moderation systems. By demonstrating that current state-of-the-art detectors plateau at F1 $\approx$ 0.89–0.94 and that 54.0% of human readers are deceived by intent-conditioned disinformation, we provide empirical grounding for the urgency of continued investment in detection research. We hope this benchmark accelerates progress toward robust, intent-aware detection systems that can protect democratic discourse and public health communication.

Human Participants. The human evaluation (N = 1,167) was conducted via Prolific under a protocol reviewed and granted exempt status by the Institutional Review Board. Participants provided informed consent, were compensated at a rate consistent with Prolific's minimum pay guidelines, and were fully debriefed upon completion. No personally identifiable information was collected or retained. Participants were not exposed to graphic, hateful, or psychologically distressing content; all articles were drawn from mainstream English-language news outlets and concern topics typical of everyday news coverage.

Data Privacy. All source articles were collected from publicly available news outlets under fair-use provisions for non-commercial research. No user data, private communications, or social media content were collected. Personally identifiable information appearing incidentally in news articles (e.g., named private individuals in court reporting) was filtered during dataset construction. The dataset does not contain medical records, financial data, or other sensitive personal information.

Limitations and Future Safeguards. We acknowledge two residual risks. First, a sufficiently motivated actor with access to the benchmark could attempt to reverse-engineer manipulation patterns from the annotation fields. We consider this risk low given the institutional access controls, but recommend that future work explore differential-privacy techniques for annotation release. Second, the current release covers only English-language content.

B Dataset Construction Details

B.1 Fact-Checking Data Sources

We aggregate fact-checked articles from 14 independent organizations spanning a 30-year period (1995–2025), yielding a corpus of 45,598 entries. Sourcing from multiple outlets—ranging from large general-purpose fact-checkers such as *Snopes* and *PolitiFact* to domain-specific organizations such as *Science Feedback*—mitigates framing biases inherent to any single source and ensures broad topical coverage across political, scientific, and social domains.

Table A1: Articles collected from each fact-checking organization.

Fact-Checking Organization	**# Articles**
Snopes	20,768
PolitiFact	9,530
Truth or Fiction	3,715
Gigafact	3,159
FactCheck.org	2,406
Lead Stories	2,403
Voice of America News	1,994
Science Feedback	916
RumorGuard	358
Check Your Fact	159
The Gazette	55
Lead Stories (regional)	52
News4Jax	42
ClickOrlando	41
Total	**45,598**

B.2 News Topic Categories

The Fake News Profiling Agent classifies each article into one of 29 topic domains, derived through a qualitative audit of the full fact-checked corpus and validated against established news taxonomy frameworks. These categories span a broad spectrum of public-interest subjects—from *Elections and Voting* and *Public Health* to *Consumer Safety* and *Foreign Policy*—ensuring that the resulting dataset captures the topical diversity observed in real-world disinformation campaigns. Using the GPT-4o-mini large language model, the 17,395 real news articles were categorized into the 29 distinct topics listed in Table A2.

Table A2: List of 29 news topic categories, arranged alphabetically.

Civil Rights and Liberties	Housing and Urban Development
Consumer Safety	Immigration
Crime and Justice	Labor and Employment
Culture and Society	Legal Issues
Drugs and Alcohol	Military and Defense
Economy and Finance	Nat. Disasters & Climate Change
Education	Public Health
Elections and Voting	Public Safety
Environment and Energy	Religion
Ethics and Transparency	Social Issues
Food and Agriculture	Sports and Recreation
Foreign Policy	Technology and Science
Government and Politics	Transportation and Infrastructure
Health and Medicine	Welfare and Social Services
History and Heritage	

B.3 Disinformation Intent Categories

We adopt a 17-subcategory intent taxonomy organized under five macro categories—*Political Agenda*, *Financial Gain*, *Social Media Attention*, *Mischievous Intent*, and *Scientific Quackery*—adapted from established literature (Jr. et al., 2018; Lazer et al., 2018; Allcott and Gentzkow, 2017; Vosoughi et al., 2018). This finer-grained scheme, compared to the 5–7 category frameworks used in prior work (Jr. et al., 2018; Zhang and Ghorbani, 2020), enables the agent to distinguish nuanced differences such as those between state-sponsored propaganda and financially motivated clickbait. We have classified the intents based on the 45,598 fact-checked articles and then using the Decaying vector (described in Section 3.3) we have assigned intents for the articles. The full list of intents is shown in Table A3.

Table A3: 17 disinformation intent subcategories mapped to 5 broad categories, ordered by frequency.

Category	**Subcategories**
Political Agenda	Electoral Influence, Policy Promotion, Ideological Advocacy, Promotion of Hate, Personal Vendettas, Ideological Manipulation
Financial Gain	Corporate Rivalry, Product Endorsements, Market Manipulation
Social Media Attention	Viral Sensation, Outrage Generation, Parody and Satire
Scientific Quackery	Health Misinformation, Environmental Misrepresentation, Technological Fallacies
Mischievous Intent	Pranks and Hoaxes, Cultural Commentary

B.4 Real News Sources

We collect 17,395 articles from ten credible news outlets selected to represent a diverse range of ideological leans, article lengths, and domain focuses—spanning national broadcast networks, cable news

channels, and progressive and centrist print outlets. This deliberate diversity ensures that the extracted writing-style profiles σ_i reflect genuine cross-outlet variation in tone, attribution conventions, and discourse structure, rather than converging on a single journalistic register.

Table A4: Distribution of real news articles used for generation, ordered by article count.

News Source	Article Count
CNN	4,245
CNBC	3,452
Mother Jones	2,256
NBC News	1,998
USA Today	1,779
ABC News	1,107
Fox News	829
Chicago Tribune	671
CBS News	582
The Nation	476
Total	**17,395**

B.5 Dataset Details

`GenFakeNewsNet` comprises 34,790 articles (17,395 real–fake pairs) spanning 10 English-language news outlets, 29 topic categories, and 17 disinformation intent subcategories. Each instance consists of a source real article paired with a style-aware, intent-conditioned fake counterpart generated by the multi-agent pipeline described in Appendix C.2.

The dataset is organised into five field groups, fully described in Tables A5 and A6: (i) **Article Content** — the original and generated article texts, masked version, and factual summary; (ii) **Semantic Annotations** — topic, disinformation intent, emotional tone, and generation goal; (iii) **Manipulation Metadata** — paragraph-level manipulation plan, operator types, and subtlety levels; (iv) **Style Profile** (σ_i) — outlet-specific stylometric features including tone, sentence length, attribution style, and discourse structure; and (v) **Quality & Integrity Verification** — detection risk score, hallucinated entity lists, fabricated quote counts, and journalist reasoning records.

Table A5: Complete field reference for GenFakeNewsNet (Part 1 of 2).

Group	Field	Type	Description
Article Content	article_id	Integer	Unique numeric identifier for the instance.
	original_real_news	String	Full text of the source real news article used as the generation seed.
	original_real_news_title	String	Title of the original real article (empty if not available from the outlet).
	fake_news_title	String	AI-generated headline for the fake article, conditioned on the article body and disinformation goal.
	rewritten_fake_news	String	Full text of the generated fake article after style-aware editing and evasion verification.
	masked_version	String	The real article with fact-central sentences replaced by [MASKED_SENTENCE] tokens; used as the primary generation prompt input.
	masked_sentences	List[String]	The exact sentences that were masked, preserving the original wording for analysis of injected disinformation.
	original_summary	String	A concise factual summary of the original real article, generated independently to serve as a ground-truth reference.
Semantic Annotations	topic	String	One of 29 news topic categories (e.g., *Sports and Recreation*, *Government and Politics*). See Appendix B.2.
	intent	String	One of 17 disinformation intent subcategories spanning 5 broad groups (e.g., *Parody and Satire*, *Electoral Influence*). See Appendix B.3.
	intent_definition	String	Natural-language definition of the assigned intent category, providing supervision signal for intent classification tasks.
	fake_news_goal	String	A detailed paragraph specifying the disinformation objective: what narrative shift to achieve, which sentences to target, and how to reframe the original article. Serves as the generation directive.
	article_emotional_tone	String	Coarse emotional register of the article (*balanced*, *alarming*, *outraged*, etc.), inferred from lexical and discourse features.
	primary_sources_detected	List[String]	Named persons or organizations cited as primary sources in the article, extracted during the journalist-framing stage.
Manipulation Metadata	manipulation_plan_summary	List[Object]	Paragraph-level log of the manipulation plan. Each entry records: paragraph index, whether it was manipulated (true/false), the operator applied (e.g., *emphasis_shift*, *juxtaposition*, *attribution_reframe*), and the subtlety level (*low*, *medium*, *high*).
	num_paragraphs_total	Integer	Total number of paragraphs in the article.
	num_paragraphs_manipulated	Integer	Number of paragraphs that received disinformation edits; supports analysis of manipulation density.
	rewrite_in_different_way	Boolean	Flag indicating whether the full article structure was reorganised (true) or manipulated in-place (false).

Table A6: Complete field reference for GenFakeNewsNet (Part 2 of 2).

Group	Field	Type	Description
Style Profile (σ_i)	writing_level	String	Outlet-level writing register (e.g., *National News*, *Investigative Journalism*).
	tone	String	Dominant article tone (e.g., *Informative and assertive*, *Neutral and analytical*).
	sentence_length_avg	String	Average sentence length as a token range (e.g., *24–30 words*).
	paragraph_structure	String	Typical paragraph composition (e.g., *mostly single-sentence paragraphs*).
	voice / tense	String	Grammatical voice and tense conventions of the outlet (e.g., *Third-person objective*, *Simple past with present context*).
	attribution_style	String	How the outlet attributes statements to sources (e.g., *Named expert quotes*).
	quotation_usage	String	Typical pattern of direct quotation (e.g., *Extensive direct quotes from named officials*).
	hedging_language	List[String]	Hedge words characteristic of the outlet (e.g., *however*, *suggests*, *reportedly*).
	narrative_flow	List[String]	Ordered list of discourse moves describing the article's progression (e.g., Announcement → Background → Expert Commentary).
	lexical_features	Object	Degree ratings for emotive-word density, institutional-term density, and sensational-phrase density.
	length_stats	Object	Article-level statistics: total words, total characters, sentence count, average and standard deviation of tokens per sentence, paragraph count, and per-paragraph sentence counts.
	lexical_stats	Object	Type-token ratio (TTR), top transition phrases, domain-specific terms, and attribution verbs.
	discourse_structure	Object	Section-level breakdown (e.g., Lede, Background, Expert Quotes) with word counts and percentage-of-article for each section.
	precise_structure	Object	Paragraph and sentence count breakdown, counts of opening/body/conclusion paragraphs, opening style, development pattern, and conclusion type.
Quality & Integrity Verification	detection_risk_score	Float [0,1]	Continuous score estimating how detectable the fake article is by automated classifiers; lower values indicate harder-to-detect disinformation (threshold calibrated on held-out data).
	detection_pass	Boolean	true if the article passed all 17 evasion checks and the risk score is below threshold, confirming sufficient realism before inclusion.
	hallucinated_entities	List[String]	Named entities introduced by the model that do not appear in the original real article; populated by the hallucination-detection check.
	fabricated_numbers	List	Numerical values fabricated during generation that were not present in the source article.
	fabricated_quotes_count	Integer	Count of direct quotes that were altered or entirely invented relative to the original.
	reframe_summary	Object	Structured record of any paragraph-level narrative reframings applied by the evasion system, including counts and affected paragraph indices.
	journalist_reasoning	Object	Per-article interpretability record covering: story angle, narrative structure, source hierarchy, attribution patterns, fact-selection strategy, tone and register, reader psychology, and paragraph-level role labels. Unique to GenFakeNewsNet.
	generation_time_seconds	Float	Wall-clock time in seconds for end-to-end generation of the article, including all LLM calls and verification passes.

C Generation Pipeline Details

C.1 Style Profile Structure

The style profile σ_i is a structured JSON object capturing both quantitative surface statistics (e.g., sentence-length distributions, type–token ratio) and qualitative attributes inferred by the Real News Profiling Agent (e.g., tone, attribution style, hedging expressions, narrative flow). The complete set of sub-fields—including their types, descriptions, and example values—is documented in Tables A5 and A6 (Group 4: *Style Profile*).

C.2 Generator Pipeline

The eight-step generator processes each real article a_i into a style-faithful fake article f_i through a combination of deterministic algorithms and targeted LLM calls. Steps 1, 4, and 7 are fully algorithmic; Steps 2, 3, 5, and 8 invoke GPT-4o-mini; and Step 6 combines algorithmic validation with LLM-driven rewriting to ensure both detection evasion and factual plausibility.

1. **Linguistic fingerprinting** (algorithmic): Extract per-paragraph fingerprints covering sentence structure, vocabulary profile, emotional tone, hedging patterns, and paragraph type.

2. **Disinformation goal** (1 LLM call): Formulate g_i from the article, its masked version, and the structural analysis.

3. **Journalist framing** (1 LLM call): Model the original journalist's narrative framing, source hierarchy, and attribution patterns.

4. **Manipulation planning** (algorithmic): Assign distortion operators to each paragraph using a dynamic manipulation ratio $r \in [0.25, 0.72]$.

5. **Paragraph cloning** ($|\mathcal{P}_i|$ LLM calls): Generate each paragraph conditioned on its fingerprint, disinformation directive, and running context.

6. **Detection & evasion** (algorithmic + targeted LLM): Run the 17-check system; reframe flagged paragraphs through up to 3 passes per paragraph and 2 article-level iterations.

7. **Formatting alignment** (algorithmic): Enforce structural and typographic fidelity to the source outlet's conventions.

8. **Headline generation** (1 LLM call): Produce headline h_i conditioned on the article body and disinformation goal g_i.

Total LLM calls per article: $3 + |\mathcal{P}_i|$ plus any reframing passes.

Fake News Generation Procedure Given a masked article m_i, an assigned intent I_i, and an outlet-specific style profile σ_i, the generation agent produces the final synthetic article f_i through paragraph-level rewriting. The pipeline contains eight stages. Step 1 extracts per-paragraph linguistic fingerprints, including sentence structure, hedging patterns, transition usage, attribution style, and source hierarchy. Step 2 formulates a disinformation goal g_i conditioned on the assigned intent and topic. Step 3 models the original journalist's framing, including source ordering and evidential emphasis. Step 4 plans distributed micro-manipulations across the article. Step 5 rewrites masked content paragraph by paragraph. Step 6 applies verification and reframing. Step 7 performs final typographic alignment. Step 8 generates a headline conditioned on the rewritten article and g_i.

The manipulation plan uses a two-tier taxonomy. The first tier adapts sentence-level propaganda techniques from prior work (Da San Martino et al., 2019) into paragraph-level operations: *tone_shift*, *emphasis_shift*, *attribution_reframe*, and *causal_reframe*. The second tier introduces article-level framing operations: *omission*, *amplification*, and *juxtaposition*. Unlike sentence-level replacement methods, these operations are composed across paragraphs and conditioned jointly on disinformation intent, topic, and outlet style.

To preserve surface form, each paragraph p_j is generated with constraints on sentence and paragraph length:

$$|s'_k| = |s_k| \pm 2, \quad |\mathbf{p}'_j| = |\mathbf{p}_j| \pm 5 \qquad (5)$$

where $|s_k|$ and $|s'_k|$ are the word counts of original and generated sentences. Additional constraints enforce identical sentence count, matched starters, reuse of original transition words where possible, preserved emotional valence, and similar punctuation patterns. Tiered anti-hallucination rules prohibit fabricated named entities, direct quotes, and statistics.

C.3 17-Check Verification System

The 17-check verification system is applied after paragraph cloning to ensure that each generated article is both sufficiently realistic to evade automated detectors and free of detectable hallucina-

tions. Checks are organized into two complementary categories: those targeting statistical signals exploited by ML-based classifiers, and those identifying factual inconsistencies or AI-generation artifacts.

ML-detector evasion checks (8): word-count alignment, sentence-length burstiness, vocabulary overlap, transition-word consistency, lexical repetition, perplexity proxy, entity consistency, formatting fidelity.

Hallucination and bias detection (9): fabricated entities, fabricated statistics, fabricated quotes, sentiment reversal, conflict downplay, ending sentiment shift, semantic redundancy, hedging density drift, AI-pattern fingerprints (LLM buzzwords, filler phrases).

When the aggregate risk score exceeds a threshold calibrated on a held-out validation set, flagged paragraphs enter an evasion reframing loop: up to 3 passes per paragraph and 2 article-level iterations.

C.4 Hyperparameters

Table A7 summarizes the key hyperparameters used across all pipeline stages, including temperature settings for each LLM call type, token budgets for goal construction and paragraph generation, rate limits applied during profiling and generation, and algorithmic parameters for KeyBERT-based sentence scoring and intent decay weighting. These values were selected empirically to balance generation quality, stylistic fidelity, and computational cost.

Table A7: Key hyperparameters used across pipeline stages.

Stage	Parameter	Value
Profiling (all)	Temperature	0.1
Style profiling	Temperature	0.3
Style profiling	Max tokens	2,048
Profiling	Rate limit	30 req/min
Generation	Rate limit	20 req/min
Goal construction	Max tokens	384–1,536
Paragraph generation	Max tokens	256–2,048
Full-article fallback	Max tokens	8,192–32,768
Manipulated paragraphs	Temperature	0.7
Non-manipulated paraph.	Temperature	0.5
Headline generation	Temperature	0.6
KeyBERT	Top keyphrases	15
KeyBERT	N-gram range	(1, 2)
Decay rate λ	Intent scoring	0.1
Sliding window w	Intent assignment	2 months

C.5 Models and Compute

Model sizes. The generation pipeline uses GPT-4o-mini as the backbone for all four agents (OpenAI API; parameter count not publicly disclosed). KeyBERT sentence embeddings are computed with `all-MiniLM-L6-v2` (22.7M parameters). Detection baselines: BOW / TF-IDF / Ngrams+SVM are non-parametric or near-zero; 3HAN ≈1.2M; Conv-FFD and ML-Transformer ≈3–7M; BERT+MLP, BERT+BiGRU, and BERT+CNN+MLP use BERT-base-uncased (110M). LLM-based detectors (GPT-4o-mini, FactAgent, ARG) are accessed through the OpenAI API. The eight open-source zero-shot LLM judges are Falcon3-7B-Instruct (7B), Llama-3.1-8B-Instruct (8B), Mistral-7B-Instruct (7B), Phi-3.5-mini-Instruct (3.8B), Qwen2.5-7B-Instruct (7B), Starling-LM-7B-beta (7B), Zephyr-7B-beta (7B), and Gemma-2-9B-it (9B).

Compute budget. End-to-end generation of the 17,395-pair corpus consumed approximately USD $450 in OpenAI API credits (GPT-4o-mini), with an average wall-clock runtime of 3–5 minutes per article including all four agents and the verification–revision loop. Zero-shot inference for the eight open-source LLM judges over the full 34,790-article evaluation set required approximately 10 hours of wall-clock time on 2×NVIDIA RTX 5090 GPUs and incurred no additional monetary cost. The traditional-ML baselines (BOW, TF-IDF, Ngrams+SVM, 3HAN) and KeyBERT extraction ran on CPU in well under one hour combined.

Infrastructure. Generation and API-based detection (GPT-4o-mini, FactAgent, ARG) ran through the OpenAI API with only CPU-side orchestration locally. Open-source LLM judging and BERT-based detector training used a workstation with 2×NVIDIA RTX 5090 (31 GB VRAM each) under CUDA 12 and PyTorch 2.x.

C.6 End-to-End Generation Walkthrough

We illustrate the Fake News Generation Agent end-to-end on a single seed article: Article #5700, a CNN report from 19 December 2022 covering world leaders' congratulations to Argentina following its 2022 FIFA World Cup victory (topic *Sports and Recreation*; assigned intent *Outrage Generation*). We chose this article because it exhibits the *distributed* manipulation pattern that distinguishes `GenFakeNewsNet`—five of eight paragraphs are rewritten under a single article-level goal, using three different operator types (including one tier-2 *document-level framing* operator)—and it does so on a non-politically-sensitive topic. The article was not selected for empirical decep-

tion success: per-article human-study labels are not released, so corpus-level deception is reported only in aggregate (§5.3, 54.0% Real-rated). The machine-side evidence available per article is included with the example: it passes the 17-check verifier on the first attempt with risk score 0.056, no hallucinated entities, no fabricated numbers, and no fabricated quotes, placing it well inside the corpus's clean-pass region. The seed has 290 words across 8 paragraphs; below we trace each of the four stages from §3.4.

Inputs. The agent receives three structured inputs from the upstream agents. **(a) Masked article** m_i: the seed article with one fact-central sentence masked by the Narrative Distortion Planning Agent (§3.3; see "Upstream masking" below). **(b) Assigned intent** I_i: *Outrage Generation*, sampled from the temporally weighted topic–intent distribution. **(c) Style profile** σ_i: an outlet-specific JSON profile from the Real News Profiling Agent; relevant fields include `writing_level`=*National News*, `tone`=*Neutral and fact-driven*, `voice`=*Third-person objective*, `attribution_style`=*Named expert quotes*, mean sentence length 24.1 tokens, and type–token ratio 0.53.

Upstream masking (from the Narrative Distortion Planning Agent). The masked article m_i is produced by the procedure in §3.3: KeyBERT (`distilbert-base-nli-mean-tokens` backbone, $(1, 2)$-gram phrases, top-K=15 keywords, English stop-word list) extracts the article's most semantically central key-phrases; each sentence s_j is then scored by Eq. 4—the count of those top-15 phrases it contains—and the top-m=1 sentence is selected for masking, subject to the constraints (i) the lede is never masked, (ii) no two masked sentences may be consecutive, and (iii) ties are broken by score. For Article #5700 the procedure selects the final sentence of the Macron-tweet paragraph (P3), "*Congratulations to Argentina for their victory*," as the highest-keyword-density non-lede sentence—it concentrates the entities *Argentina*, *victory*, and *congratulations*, which all appear among the article-level top phrases for this seed. The released dataset stores the resulting masked sentence in the `masked_sentences` field and the placeholder-substituted article in `masked_version`; the exact keyword set is not retained on disk but is reproducible from the seed text by running KeyBERT with the configuration above.

Stage 1 (Linguistic fingerprinting). The agent extracts the article's structural fingerprint: 8 paragraphs, mean 1.5 sentences/paragraph, emotional tone *critical*, and 4 named primary sources (*Joe Biden*, *Vladimir Putin*, *Emmanuel Macron*, *Alberto Fernández*). A journalist-reasoning sub-call records the article's story angle ("world leaders responding to Argentina's World Cup victory") and narrative structure ("inverted pyramid with stacked attributed quotes"), which become reproduction targets for the rewrite.

Stage 2 (Goal formulation). Conditioned on I_i=*Outrage Generation* and the Stage 1 fingerprint, the agent emits a structured disinformation goal g_i in a single LLM call. g_i contains directives at two levels simultaneously—a *macro* (article-level) objective and *micro* (per-paragraph and per-mask) directives—so that the per-paragraph instructions are jointly consistent with the article-level objective rather than independently authored. All layers are stored together in the dataset's `fake_news_goal` field.

Macro objective (article-level).

> *"Instill a belief that Argentina's World Cup victory was not just a sporting achievement but a symbol of geopolitical tension, suggesting that world leaders' congratulations were veiled attempts to distract from their countries' own issues . . . framing Argentina's victory as a moment of superficial unity rather than genuine celebration."*

Micro directive 1 (per masked sentence). For the one masked sentence selected upstream (P3's "Congratulations to Argentina for their victory"), g_i specifies the intended replacement:

> *"Replace with a statement emphasizing that these congratulations were a façade concealing deeper tensions, such as: 'While congratulations were extended to Argentina, they served as a mere distraction from the turmoil facing these world leaders at home.' "*

Micro directive 2 (per individual paragraph). For *each* non-masked paragraph, g_i specifies how that specific paragraph should be shifted to support the macro objective. The directives for Article #5700 are: *P0 (lede):* frame the leaders' congratulatory statements as obligatory rather than heartfelt; *P1 (context):* introduce the idea that the victory is being exploited for political gain; *P2–P5 (quote-heavy):* retain the original direct quotes verbatim but recontextualize them as political posturing; *P6 (narrative):* shift focus toward the victory

being perceived internationally as a smokescreen; *P7 (concluding):* cast doubt on what the celebrations mask about the political landscape. These per-paragraph directives are accompanied by per-content-type rules that apply across paragraphs: criticism is softened, attributed quotes are reframed without altering quoted text, data is recontextualized, and narrative passages emphasize skepticism about leaders' intentions.

The macro + per-paragraph + per-mask layers are what the Stage-3 planner reads to convert into a concrete per-paragraph operator plan: the macro objective sets the overall frame, the per-mask directive determines what surface form the masked-sentence rewrite should take, and each per-paragraph directive maps directly onto an operator choice for that paragraph (e.g. "frame as obligatory rather than heartfelt" on P0 → `emphasis_shift`; "recontextualize quotes as political posturing" on P2, P4, P5 → `attribution_reframe`; the P3 quote-retention-plus-recontextualization directive → tier-2 `juxtaposition`).

Stage 3 (Manipulation planning). The planner assigns operators from the two-tier taxonomy (§3.4) to each paragraph, distributing distortions across the article. The resulting per-paragraph plan is:

P#	Operator (tier)	Level
0	`emphasis_shift` (T1)	medium
1	—	—
2	`attribution_reframe` (T1)	low
3	`juxtaposition` (T2)	medium
4	`attribution_reframe` (T1)	medium
5	`attribution_reframe` (T1)	low
6	—	—
7	—	—

Five of eight paragraphs are manipulated under one coordinated plan; three (P1, P6, P7) are passed through unchanged. The plan mixes a tier-1 rhetorical operator (`emphasis_shift`) with three tier-1 attribution-verb downgrades and one tier-2 document-level framing operator (`juxtaposition`), so the manipulation is genuinely distributed rather than concentrated in a single paragraph.

Stage 4 (Paragraph rewriting under constraints). Each manipulated paragraph is regenerated conditioned on its Stage-1 fingerprint, its assigned operator from Stage 3, and the goal g_i, under the surface-form constraints in §3.4. Three representative rewrites demonstrate how the operators compose into a coordinated narrative shift:

P0 (`emphasis_shift`, medium). *Original:* "United States President Joe Biden, Russian President Vladimir Putin and French President Emmanuel Macron are amongst world leaders who have *congratulated* Argentina President Alberto Fernández . . . "

Rewritten: "*While congratulations were extended to Argentina, they served as a mere distraction from the turmoil facing these world leaders at home,* including United States President Joe Biden, Russian President Vladimir Putin, and French President Emmanuel Macron . . . "

The lede reframes the entire article: the original presents congratulations as the news; the rewrite presents them as a cover for "turmoil" faced by the same named leaders, directly advancing g_i. All four named entities are preserved.

P2 (`attribution_reframe`, low). *Original:* ". . . President Biden <u>wrote</u> on Twitter Monday . . . " *Rewritten:* ". . . President Biden <u>claimed</u> on Twitter Monday . . . "

The neutral reporting verb *wrote* is replaced with the skeptical *claimed*, casting the same quote in doubt. The same operator is applied consistently to subsequent attribution verbs across the article: P4 changes "Russia's Ministry of Foreign Affairs *said*" to "*claimed*", and P5 changes "President Fernández *confirmed*" to "*acknowledged*". Individually each substitution is minimal; collectively they produce a coordinated tone of doubt across every attributed quote in the article.

P3 (`juxtaposition`, medium). The Macron quote (verbatim from the seed) is retained in place but the surrounding paragraph is repositioned so that the leader's congratulatory tweet is juxtaposed against the framing established in P0; the operator inserts no new claim but recontextualizes an unchanged quote, which is the defining property of the tier-2 `juxtaposition` operator.

Stage 4.5 (Verification). The 17-check verifier scores the assembled article. For this example it reports an overall risk score of 0.056 (well below the reframe threshold), 0 hallucinated entities, 0 fabricated numbers, and 0 fabricated quotes. The article passes on the first attempt; no paragraphs enter the reframe loop.

Final output. The rewritten article preserves the original 8-paragraph structure and every named

entity (Biden, Putin, Macron, Fernández, Messi) and every direct quote, while advancing the geopolitical-distraction narrative specified by g_i through a coordinated combination of lede framing, attribution-verb downgrading across multiple paragraphs, and one tier-2 juxtaposition. The generated headline—"World Leaders' Congratulations to Argentina's Victory Mask Political Turmoil at Home"—is conditioned on the rewritten body and g_i in Stage 7 of the generator (see Appendix C.2).

C.7 Open-Source LLM Generation Failures

To justify our choice of GPT-4o-mini as the sole generation backbone, we document the two failure modes that disqualified open-source alternatives. Figure A1 shows outputs from LLaMA-3-8B and Gemma-7B on the USA TODAY health-nutrition seed article (oatmeal benefits) used in our qualitative comparison (§E), under identical prompting conditions. LLaMA produces degenerate repetition—recycling the same sentence fragment across consecutive paragraphs until the context window is exhausted. Gemma fabricates entities and events with no grounding in the source article, rendering the output unusable as a disinformation benchmark entry. GPT-4o-mini produces neither failure mode, preserving source structure while executing the intended pragmatic reframe.

• ORIGINAL ARTICLE (USA TODAY, health-nutrition seed)

You've likely heard health and nutrition experts highlight the benefits of eating a healthy, well-balanced breakfast. Could oatmeal be the answer? "Oats are a super nutrient-dense food that can support digestion, heart health and even weight management," registered dietitian Miranda Galati tells USA TODAY. "They contain a unique beta-glucan fiber that's particularly powerful for lowering cholesterol and improving blood sugar." **[...]** Galati recommends pairing them with other nutritious ingredients to make them "more filling and blood sugar friendly."

• LLaMA-3-8B FAIL: DEGENERATE REPETITION

You've likely heard that breakfast may be overrated. Oats are often cited as a nutritious food, but oats are often cited as a nutritious food that may not suit everyone. Miranda Galati, a registered dietitian, tells USA TODAY that oats are often cited as a nutritious food, but oats are often cited as a nutritious food that may not suit everyone, as oats are often cited as a nutritious food. "They contain beta-glucan fiber," though oats are often cited as a nutritious food that may not suit everyone, and oats are often cited as a nutritious food that may not suit everyone. Galati recommends pairing oats, which are often cited as a nutritious food, with other ingredients, as oats are often cited as a nutritious food that may not suit everyone.

• Gemma-7B FAIL: HALLUCINATION

Health experts have long debated the risks of oatmeal, particularly following a landmark 2019 WHO report that classified beta-glucan fiber as a potential endocrine disruptor. Registered dietitian Miranda Galati, speaking at the International Nutrition Summit in Geneva last April, warned that oats "can trigger systemic inflammation in a majority of adults." NASA researchers studying dietary habits aboard the ISS have also flagged oatmeal as incompatible with cardiovascular health in low-gravity environments, raising fresh concerns about its suitability as a breakfast food. A class-action lawsuit filed against Quaker Oats in 2023 by 4,200 plaintiffs further underscores the growing skepticism around beta-glucan supplementation. Galati now advises patients to consult the newly established Federal Oat Safety Commission before incorporating oats into their daily diet.

• GENFAKENEWSNET (OURS) BEST: STRATEGIC REFRAMING

You've likely heard that breakfast may be overrated for some individuals — rising interest in oatmeal is based on marketing claims rather than solid health benefits for everyone. [...] Miranda Galati tells USA TODAY: "Oats are a super nutrient-dense food that may support digestion... although some individuals experience adverse reactions to oats." **[...]** For those monitoring blood sugar, opting for groats or steel-cut oats could still lead to the same adverse reactions... Oats can be a common allergen... pairing them with ingredients might lead to unwanted digestive issues, which raises questions about their reputation.

Degenerate repetition | Hallucinated entity / event | Framing shift | Amplified caveat

Figure A1: Generation failure modes on the USA TODAY health-nutrition seed article under identical prompting. LLaMA-3-8B degenerates into looping repetition of the same phrase fragment. Gemma-7B fabricates non-existent sources, events, and institutions with no grounding in the source article. GenFakeNewsNet, with GPT-4o-mini as its backbone, executes the intended pragmatic reframe without either failure mode, preserving source structure and named sources while shifting narrative stance.

D Evaluation Details

D.1 Zero-Shot Open-Source LLM Judges

Table A8 reports the full model-level zero-shot results summarized in the main paper. All models are evaluated on the full balanced GenFakeNewsNet dataset containing 17,395 real and 17,395 generated fake articles (34,790 articles total). We report precision, recall, and F1 on the fake class because the key question is whether open-source LLM judges can identify generated fake articles without task-specific training or in-context examples.

Model	$\mathbf{P}_{fake}$	$\mathbf{R}_{fake}$	$\mathbf{F1}_{fake}$
Falcon3-7B-Instruct	0.591	0.025	0.047
Granite-3.1-8B-Instruct	0.619	0.098	0.169
Llama-3.1-8B-Instruct	0.669	0.046	0.086
Mistral-7B-Instruct	0.630	0.043	0.081
Phi-3.5-mini-Instruct	0.659	0.009	0.017
Qwen2.5-7B-Instruct	0.695	0.070	0.127
Starling-LM-7B-beta	0.691	0.067	0.122
Zephyr-7B-beta	0.581	0.004	0.007

Table A8: Zero-shot open-source LLM detection on the fake class of the full GenFakeNewsNet dataset (17,395 real and 17,395 generated fake articles).

The results show that fake-class recall remains low across all eight judges (0.004–0.098). Thus, even when precision is moderate (0.58–0.70), the models flag only a small subset of generated fake articles as fake and misclassify most generated fake articles as real.

D.2 Supplementary Benchmark Datasets

Tables A9 and A10 report detection performance across seven supplementary benchmark datasets not included in the main evaluation: Med-MMHL, F3 Articles, F3 Social, ISOT, TruthSeeker, KDNuggets, and AdvFake. Results cover three families of methods—Traditional ML, Deep Neural Networks, and LLM-Based—and are presented as precision (P), recall (R), and F1, providing a comprehensive view of cross-dataset generalization.

Table A9: Supplementary detection results on Med-MMHL, F3 Articles, F3 Social, and ISOT.

Method	Med-MMHL			F3 Art.			F3 Soc.			ISOT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Traditional ML												
BOW (Singhania et al., 2017)	0.97	0.97	0.97	0.98	0.98	0.98	0.88	0.84	0.86	0.97	0.96	0.96
BOW+TFIDF (Singhania et al., 2017)	0.97	0.98	0.98	0.98	0.98	0.98	0.88	0.86	0.87	0.97	0.97	0.97
Ngrams+TFIDF (Singhania et al., 2017)	0.97	0.97	0.97	0.99	0.99	0.99	0.87	0.86	0.87	0.97	0.97	0.97
Ngrams+SVM (Singhania et al., 2017)	0.98	0.97	0.98	0.99	0.99	0.99	0.87	0.89	0.88	0.97	0.97	0.97
Deep Neural Networks												
BERT+MLP (Zhang et al., 2023b)	0.98	0.99	0.98	0.99	0.99	0.99	0.91	0.90	0.91	0.99	0.99	0.99
BERT+BiGRU (Zhang et al., 2023b)	0.98	0.98	0.98	0.99	0.99	0.99	0.91	0.91	0.91	0.99	0.99	0.99
BERT+CNN+MLP (Zhang et al., 2023b)	0.98	0.98	0.98	0.99	0.99	0.99	0.91	0.91	0.91	0.99	0.99	0.99
Conv-FFD (Zhang et al., 2023a)	0.97	0.97	0.97	0.99	0.99	0.99	0.89	0.90	0.89	0.98	0.99	0.98
ML-Transformer (Zhang et al., 2023a)	0.97	0.98	0.97	0.99	0.99	0.99	0.89	0.89	0.89	0.99	0.99	0.99
3HAN (Singhania et al., 2017)	0.58	0.85	0.69	0.57	1.00	0.73	0.56	0.97	0.71	0.57	1.00	0.73
LLM-Based												
ARG (Hu et al., 2024)	0.94	0.95	0.94	0.93	0.93	0.93	0.92	0.91	0.91	0.95	0.94	0.94
Zero-shot (Li et al., 2024)	0.82	0.65	0.73	0.85	0.68	0.76	0.80	0.62	0.70	0.84	0.64	0.73
FactAgent (Li et al., 2024)	0.88	0.82	0.85	0.89	0.80	0.84	0.85	0.78	0.81	0.90	0.81	0.85

Table A10: Supplementary detection results on TruthSeeker, KDNuggets, and AdvFake.

Method	TruthSeeker			KDNuggets			AdvFake		
	P	R	F1	P	R	F1	P	R	F1
Traditional ML									
BOW (Singhania et al., 2017)	0.95	0.96	0.96	0.99	0.99	0.99	0.49	0.50	0.50
BOW+TFIDF (Singhania et al., 2017)	0.96	0.97	0.96	0.99	0.99	0.99	0.49	0.49	0.49
Ngrams+TFIDF (Singhania et al., 2017)	0.95	0.96	0.96	0.99	0.99	0.99	0.48	0.49	0.49
Ngrams+SVM (Singhania et al., 2017)	0.96	0.97	0.97	0.99	0.99	0.99	0.51	0.30	0.38
Deep Neural Networks									
BERT+MLP (Zhang et al., 2023b)	0.97	0.98	0.97	0.99	0.99	0.99	0.58	0.42	0.49
BERT+BiGRU (Zhang et al., 2023b)	0.97	0.97	0.97	0.99	0.99	0.99	0.59	0.61	0.60
BERT+CNN+MLP (Zhang et al., 2023b)	0.97	0.97	0.97	0.99	0.99	0.99	0.52	0.63	0.57
Conv-FFD (Zhang et al., 2023a)	0.96	0.96	0.96	0.99	0.99	0.99	0.53	0.70	0.60
ML-Transformer (Zhang et al., 2023a)	0.96	0.97	0.96	0.99	0.99	0.99	0.56	0.29	0.38
3HAN (Singhania et al., 2017)	0.57	1.00	0.72	0.57	1.00	0.73	0.47	0.22	0.30
LLM-Based									
ARG (Hu et al., 2024)	0.92	0.93	0.92	0.94	0.94	0.94	0.91	0.92	0.91
Zero-shot (Li et al., 2024)	0.82	0.65	0.73	0.83	0.64	0.72	0.82	0.71	0.76
FactAgent (Li et al., 2024)	0.88	0.82	0.85	0.89	0.82	0.85	0.86	0.84	0.85

D.3 Human Evaluation Details

Study design. Participants were recruited via Prolific and compensated at $1 per participant; each participant took approximately 5 minutes on average to complete the article evaluation. Each participant read one article (randomly assigned from GenFakeNewsNet, VLPFN, or PropaNews) and answered four questions (Q1–Q4), whose full wording and answer choices are given in Table A11. Briefly, Q1 asked whether the article was *Real* or *Fake*; Q2 asked fake-labeling readers to identify the disinformation sub-intent from a 5-category, 17-subcategory taxonomy; Q3 asked those same readers to select all cues that gave the article away (multi-select, 8 options); and Q4 asked all readers to rate the article's logical soundness on a 5-point scale (*Not Sound* to *Very Sound*). Reading time was recorded automatically by the Prolific platform.

Table A11: Human evaluation survey questions and answer choices. Q2 and Q3 are conditional on a *Fake* response to Q1. Q3 is multi-select; all other questions are single-select.

Q	Question	Type	Answer Choices
Q1	Please read the article carefully and decide whether you believe it is `Fake` (contains misinformation / disinformation) or `Real` (trustworthy and factually accurate).	Single-select	*Real*; *Fake*
Q2	If you think the article is fake, please select the specific sub-intent you believe was behind creating the article.	Single-select(if Q1 = Fake)	**Political Agenda:** Electoral Influence, Policy Promotion, Ideological Advocacy, Ideological Manipulation, Promotion of Hate, Personal Vendettas **Financial Gain:** Corporate Rivalry, Product Endorsements, Market Manipulation **Social Media Attention:** Viral Sensation, Outrage Generation, Parody and Satire **Mischievous Intent:** Pranks and Hoaxes, Cultural Commentary **Scientific Quackery:** Health Disinformation, Environmental Misrepresentation, Technological Fallacies **Other** (free text)
Q3	If you think the article is fake, what made you think so?	Multi-select(if Q1 = Fake)	(1) Weak or illogical structure; (2) Ambiguity in the news (unclear details, vague claims); (3) Lack of flow (does not read smoothly); (4) Factually incorrect statements; (5) Poor grammar; (6) Contradictions in the news (self-conflicting statements); (7) Repetitive words, phrases, or comparisons; (8) None of the above
Q4	How logically sound is the article's reasoning?	5-point scale(all readers)	*Not Sound* (1); *Weak* (2); *Average* (3); *Strong* (4); *Very Sound* (5)

Original vs. perceived intent. Table A12 shows the intent confusion matrix for fake-labeled GenFakeNewsNet responses with a recoverable original intent label (n=137). Overall intent-matching accuracy is 55.5%. *Electoral Influence* is the clearest case (87.0%), and *Ideological Advocacy* is correctly assigned to Political Agenda in 100% of cases (though n=2). *Policy Promotion* is most ambiguous (≈39% match), with many readers perceiving it as social-media sensationalism instead.

Table A12: Original vs. perceived intent for fake-labeled GenFakeNewsNet readers (n=137). Each cell shows count (row %). **Bold*** = expected match on the diagonal. Overall intent-matching accuracy: 55.5%.

Original Intent	Polit.	Fin.	Soc. Med.	Misch.	Sci. Q.	Other
Viral Sensation (70)	11 (16)	7 (10)	**39 (56)***	9 (13)	3 (4)	1 (1)
Policy Promotion (33)	**13 (39)***	4 (12)	10 (30)	1 (3)	3 (9)	2 (6)
Electoral Influence (23)	**20 (87)***	0 (0)	2 (9)	1 (4)	0 (0)	0 (0)
Health Misinfo. † (9)	0 (0)	1 (11)	4 (44)	2 (22)	**2 (22)***	0 (0)
Ideological Adv. † (2)	**2 (100)***	0	0	0	0	0

† n<15; shown for completeness only.

Detection cues and perceived intent. Tables A13 and A14 report the aggregate detection-cue and perceived-intent distributions discussed in Section 5.3, broken down across all three generation methods.

Table A13 summarises which surface features prompted readers to label an article as fake (Q3, multi-select). PropaNews is dominated by structural and mechanical failures: "Lack of flow" (66.8%), "Weak/Illogical structure" (63.8%), and "Poor grammar" (38.4%) together account for the bulk of its detections, indicating that readers identify it through fluency and coherence breakdowns rather than factual scrutiny. VLPFN exhibits a flatter profile, with "Ambiguous claims" (49.7%) as the most frequent cue—suggesting it is flagged primarily for vagueness rather than overt errors. GenFakeNewsNet shifts the cue distribution toward higher-order concerns: "Weak/Illogical structure" (42.8%) and "Ambiguous claims" (41.1%) lead, while "Poor grammar" drops to 15.6% (less than half of PropaNews's rate). This pattern is consistent with our design goal: detections of our output reflect strategic vagueness typical of real-world propaganda, not technical generation artifacts.

Table A13: Detection cues (Q3) among fake-labeled readers (% = count / n_{fake}). Multiple cues could be selected. Bold entries mark the top cue per method.

Cue	Ours (n=180)	VLPFN (n=191)	PropaNews (n=307)
Ambiguous claims	41.1%	**49.7%**	48.5%
Weak/Illogical structure	**42.8%**	36.1%	63.8%
Lack of flow	36.7%	26.7%	**66.8%**
Factually incorrect	31.1%	28.8%	39.4%
Repetitive words/phrases	22.2%	13.6%	18.2%
Contradictions	20.0%	16.8%	21.2%
Poor grammar	15.6%	4.7%	38.4%
None of the above	2.2%	7.3%	1.6%

Table A14 reports the disinformation motive readers attributed to each fake-labeled article (Q2, single-select). "Social Media Attention" emerges as the default attribution across all three methods (39.1–41.8%), suggesting that readers treat virality and engagement-seeking as the prototypical motive for online disinformation. "Political Agenda" is the second-most-common attribution, with VLPFN attracting the highest share (35.6%) and PropaNews the lowest (26.1%). Two contrasts are notable. First, GenFakeNewsNet attracts more than twice as many "Financial Gain" attributions (9.5%) as PropaNews (3.6%), indicating that our goal-directed synthesis successfully embeds a sense of tangible commercial purpose into articles where the intent calls for it. Second, PropaNews receives a notably higher "Mischievous Intent" rate (17.0%) than either our method (9.5%) or VLPFN (8.0%)—consistent with the qualitative observation that PropaNews articles often read as implausible or prank-like rather than as targeted disinformation. Combined with the 55.5% overall intent-matching accuracy reported in the confusion matrix (Table A12), these distributions confirm that our multi-agent pipeline communicates specific, strategic intents to human readers rather than defaulting to a single recognisable motive.

Table A14: Perceived intent (Q2) among fake-labeled readers (% = count / n_{fake}).

Perceived Intent	Ours	VLPFN	PropaNews
Social Media Attention	39.1%	41.5%	41.8%
Political Agenda	31.8%	35.6%	26.1%
Mischievous Intent	9.5%	8.0%	17.0%
Financial Gain	9.5%	5.9%	3.6%
Scientific Quackery	6.1%	5.9%	7.5%
Other	3.9%	3.2%	3.9%

D.4 Quality-of-Deception Statistical Tests

This section provides the statistical tests underlying the human-evaluation claims in Section 5.3. The tables address three questions: (i) whether deception-rate differences between generation methods are statistically reliable, (ii) which article-level features are associated with being perceived as real or fake, and (iii) which factors independently predict successful deception after controlling for method and participant attributes.

Table A15 compares pairwise deception rates across methods. `GenFakeNewsNet` is significantly more deceptive than PropaNews (52.7% vs. 18.9%, $p<0.001$), while its deception rate is statistically indistinguishable from VLPFN (52.7% vs. 50.2%, ns). This supports our claim that PropaNews is more transparent to human readers, whereas our method produces articles that are perceived as real at a substantially higher rate.

Table A15: Pairwise t-tests on deception rates ($^{***}p<0.001$; ns = not significant).

Comparison	A	B	t	p
Ours–PropaNews	52.7%	18.9%	8.41	$< 0.001^{***}$
Ours–VLPFN	52.7%	50.2%	0.36	0.72 ns
PropaNews–VLPFN	18.9%	50.2%	-8.08	$< 0.001^{***}$

Table A16 examines whether surface article features explain real/fake perception. Reading time, average word length, newline count, and question count differ significantly between articles perceived as real and fake, whereas word count, capitalization ratio, and demographic variables do not. These results suggest that deception is not driven by participant demographics or simple length effects, but by how naturally and fluently the article is read.

Table A16: t-tests on article features (Real vs. Fake perception). $^{***}p<0.001$; $^{*}p<0.05$; ns = not significant.

Feature	t	p	Sig.
Reading time	7.02	< 0.001	***
Avg. word length	-4.35	< 0.001	***
Newline count	2.47	0.014	*
Question count	1.98	0.048	*
Word count	1.21	0.226	ns
Caps ratio	-0.40	0.688	ns
Age / Education / Race	<1	>0.60	ns

Finally, Table A17 reports a logistic regression predicting successful deception while controlling for article features, method, and participant demographics. Reasoning quality is the strongest positive predictor: articles judged as logically sound are much more likely to be labeled real. PropaNews has significantly lower deception odds than `GenFakeNewsNet`, while VLPFN is not significantly different from our method. This confirms that our advantage over PropaNews remains even after accounting for surface features and demographic controls.

Table A17: Logistic regression predicting successful deception (deceived = 1). Pseudo R^2=0.488; AIC = 593.7; n=816.

Predictor	β	z	Sig.
Reasoning quality (Q4)	$+2.175$	13.39	***
Reading time	-0.000005	-3.76	***
Word count	$+0.003$	2.17	*
PropaNews (vs. Ours)	-0.620	-2.10	*
VLPFN (vs. Ours)	-0.049	-0.16	ns
Gender	-0.216	-0.99	ns
Age / Education / Race	≈ 0	<1	ns

D.5 Critical-Sentence Selection Ablation

This appendix expands the ablation reported in §5.5. We compare seven sentence-selection strategies for the Critical Sentence Masking step (§3.3) on the full `GenFakeNewsNet` corpus (N=17,395 articles, mean 61.8 sentences per article).

Methods. We compare four selection strategies. *Keyword-overlap* (ours) dynamically scores each sentence by the importance scores of KeyBERT contextual keywords it contains. *Embedding-centroid* selects sentences whose contextual embeddings have highest cosine similarity to the article's mean embedding. *Lead* selects the article's leading sentences. *Random* selects sentences uniformly at random. We use a dynamic masking ratio ($\approx 4.6\%$) based on the number of sentences.

Metrics. *Factuality Gain* (FG) is the difference between the mean factuality score of the *selected* sentences and the article-wide mean factuality score; higher values indicate that the selected sentence carries more factual content than a random article sentence. *Reference Coverage@τ* (RefCov) is the fraction of selected sentences whose cosine similarity to any reference fact-central sentence (derived from independent factuality annotations) is at least τ=0.8; this measures precision against an external fact-central reference. *Reference Any-Hit@τ* (RefHit) is the binary article-level analogue: 1 if any selected sentence hits a reference at τ, else 0.

Table A18: Critical-sentence selection ablation on the full GenFakeNewsNet corpus (N=17,395). Best in **bold**.

Method	Mask Ratio	FG (↑)	RefCov@0.8 (↑)	RefHit@0.8 (↑)
Dynamic selection (masking ratio ≈ 4.6%)				
keyword-overlap (ours)	0.046	**2.624**	**0.030**	**0.030**
embedding-centroid	0.046	**2.624**	**0.030**	**0.030**
lead	0.046	1.096	0.006	0.006
random	0.046	−0.057	0.043	0.043

Findings. (i) At the dynamic configuration, *keyword-overlap* and *embedding-centroid* produce identical selections on every article, yielding identical FG, RefCov, and RefHit. Inspecting per-article outputs confirms that the two rankings converge on the same sentences for 17,395/17,395 articles, indicating that for news-article distributions of the lengths we work with, token-overlap weighted by importance is an adequate proxy for embedding-centroid distance. (ii) Both ranked scorers strongly outperform the *lead* baseline (FG 1.096) and the *random* baseline (FG -0.057), confirming that the keyword-overlap signal is informative beyond positional or chance-level selection. (iii) On the basis of (i)–(ii), we adopt KeyBERT keyword-overlap with a dynamic threshold in the released pipeline: it matches the best embedding-based method on every metric, requires no additional embedding inference, and is interpretable.

E Qualitative Generation Comparison

Figure A2 compares disinformation generation methods on 2024-election coverage, using one actual output from each method's released dataset. The GenFakeNewsNet pane is a real–fake pair from a USA TODAY article covering the September 2024 Harris–Trump presidential debate; the PropaNews and VLPFN panes are representative failure modes from their public outputs on Trump-related news. PropaNews abandons the source entirely, fabricating non-existent companies ("Diffoda Autosports", "Diageo Building") and incoherent events. VLPFN preserves journalistic surface form but invents a Pelosi–Trump exchange wholesale, producing a quotable factual inversion. In contrast, GenFakeNewsNet keeps every named entity and direct quote from the source, applying surgical framing shifts ("her lack of substantial responses raised questions about her preparedness") that invert the article's pragmatic stance without leaving statistical or factual fingerprints. Highlights denote: framing shifts, amplified content, factual contradictions, and hallucinated content.

• **ORIGINAL ARTICLE (USA TODAY, Harris–Trump debate)**

Vice President Kamala Harris and former President Donald Trump met on stage in Philadelphia for the first time in what is currently the only scheduled debate between the 2024 presidential candidates. With the election less than 60 days away and likely to be determined by razor-thin margins, expectations were high for both campaigns. **[...]** Trump, who was corrected by ABC's moderators David Muir and Linsey Davis on issues such as claims of abortion policies that allowed babies to be killed after their birth and Haitian immigrants eating pets in Springfield, Ohio, appeared rattled at times. **[...]** The evening was capped off by a key endorsement with Taylor Swift announcing her support for Kamala Harris shortly after the debate ended.

• **PROPANEWS** FAILED: TOPIC DRIFT & FABRICATION

If you haven't heard, Tesla Motors now has a weirdly popular website called "FreeTrump." The site asked its users to participate in a Google+ Hangout on Tesla's Model S 5S, but its delivery drivers actually did so on a manual basis. FreeTrump founder Howard Dinsmore had plans to create a free local-only model for the city of Bend, Oregon, by merging Tesla's two notoriously cryptic engines into a city design firm called Diageo Building. [...] The company quit funding a startup called Defus, run by right-winger Evan Conway, before the company sold to the Transportation & Infrastructure Institute for nearly $1 million.

• **VLPFN** LOW: FACTUAL INVERSION

AUSTIN, Texas — Former House Speaker Nancy Pelosi expressed confidence that former President Trump will attend the upcoming presidential debate, attributing his decision to "bravery." During her appearance at the Texas Tribune Festival, she engaged in a lively discussion with journalist Kara Swisher about her unexpected admiration for Trump. **[...]** "I know bravery when I see it," Pelosi replied. **[...]** In a surprising twist, the Trump campaign responded positively to Pelosi's comments, acknowledging her credibility.

• **SALF** MODERATE: STYLISTIC COLLAPSE

A noteworthy political encounter unfolded in Philadelphia, where Vice President Kamala Harris and former President Donald Trump engaged in what observers described as a measured exchange of policy positions. **[...]** Analysts noted that the discussion on immigration and reproductive rights appeared to reframe earlier campaign narratives, with both candidates seeking to recalibrate their public posture ahead of a tightening race. **[...]** The evening drew further attention through a high-profile cultural endorsement that some commentators viewed as redirecting the broader media cycle.

• **GENFAKENEWSNET (OURS)** BEST: STRATEGIC REFRAMING

Vice President Kamala Harris faced off against former President Donald Trump in Philadelphia, where her lack of substantial responses raised questions about her preparedness for the election, marking the only scheduled debate between the 2024 presidential candidates. With the election less than 60 days away and likely to hinge on razor-thin margins, expectations were tempered for both campaigns. **[...]** Trump, who faced corrections from ABC's moderators David Muir and Linsey Davis on matters like assertions about abortion policies permitting the killing of infants after birth and claims about Haitian immigrants consuming pets in Springfield, Ohio, seemed unsettled at moments. **[...]** The night concluded with an important endorsement as Taylor Swift declared her backing for Kamala Harris soon after the debate wrapped up.

Emphasis / framing shift | Amplified / added content | Factual contradiction | Hallucination

Figure A2: Document-level comparison of disinformation quality on 2024-election coverage. The original article and the `GenFakeNewsNet` fake are an actual real–fake pair from a USA TODAY article covering the September 2024 Harris–Trump presidential debate; the PropaNews and VLPFN samples are real outputs drawn from their released datasets on the same political-news cycle, and the SALF pane shows the adversarial-restructuring failure mode applied to the same seed. Intermediate text is omitted **[...]** for brevity. **GenFakeNewsNet** (Ours) preserves every named entity (Harris, Trump, Muir, Davis, Taylor Swift) and embedded quote while executing a surgical pragmatic reframing. Baselines fail via thematic drift and entity fabrication, factual inversion, or stylistic restructuring—each leaving statistical fingerprints that current detectors can still exploit.