

Peer Review

Review of: "Mitigating GenAI-powered Evidence Pollution for Out-of-Context Multimodal Misinformation Detection"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper moves the GenAI threat for out-of-context (OOC) image-caption misinformation from the claim to the evidence. Instead of rewriting the false caption in a neutral style, the attacker plants generated material in the pool of web evidence the detector retrieves. The authors build this pollution at scale: GPT-4 writes entity descriptions and supporting or refuting texts conditioned on the caption (§3.2, two zero-shot instructions quoted verbatim), and depth-conditional Stable Diffusion 2 produces variants of each evidence image. Table 1 gives the resulting corpus, for example, 903,067 generated texts next to 689,995 clean ones in the NewsCLIPPings training split. Four detectors (CCN, RED-DOT, SNIFFER, and zero-shot GPT-4o) are then run on clean and polluted evidence, and the abstract reports a drop of more than 9 percentage points.

The second half proposes two plug-in defenses that need no retraining. Cross-modal evidence reranking uses CLIP to rank textual evidence by the claim image and visual evidence by the claim caption, keeping the top-k (Appendix B: top-1 sentence, top-5 images, CLIP ViT-L/14). Cross-modal claim-evidence reasoning adds a consistency check between the claim image and the best-ranked textual evidence. Combined, they lift SNIFFER to 88.82% accuracy on NewsCLIPPings (+12.40) and 75.44% on VERITE (+11.15), according to §4.3.

I think the core idea is genuinely useful. Retrieval-augmented fact-checkers do assume that what the search engine returns is honest, and this paper makes that assumption testable with a large multimodal pollution set, two benchmarks, a human study, and a ratio sweep. The defense is cheap and model-agnostic, which matters for anyone running such a system in production.

2. Whether the conclusion holds on the data

Several things here are done well. The threat model is concrete, and the generation prompts are given word for word in §3.2. The authors test a pollution-ratio curve (Figure 5a, with 25% pollution costing SNIFFER 7.63 points), a second generator family (Figure 5b, DALL-E), and three alternative defenses (extra detector, vigilant prompting, reader ensemble) in Table 4. Appendix F quantifies what the reranker actually does, which many defense papers skip. That is a solid experimental skeleton.

My main difficulty is that I could not check the headline numbers myself. As rendered on Qeios, Tables 2, 3, and 4 are not legible: only their captions survive, so the "more than 9 percentage points" drop, the 35.67-point F1 fall for CCN on VERITE false claims (§4.2), and the per-detector gains exist for me only as sentences in the text. What follows is therefore built from the numbers quoted in prose and from Tables 1 and 5.

Table numbering. §4.2 says "Table 3 shows the OOC detection performance across different evidence modalities," but that is the caption of Table 2. §4.3 then says "Table 4 shows the performance ... when we incorporate the proposed defense strategies," which is the caption of Table 3. The same paragraph uses Table 4 again for the comparison with other strategies, which does match its caption. So the first two references seem to be shifted by one.

An internal inconsistency on VERITE. I checked the arithmetic of the SNIFFER gains. On NewsCLIPPings, the figures agree: $88.82 - 12.40 = 76.42$ and $87.68 - 11.26 = 76.42$, so the polluted baseline is 76.42% in both statements. On VERITE, they do not. §4.3 point 1 gives 75.44% (+11.15), implying a baseline of $75.44 - 11.15 = 64.29$. The next paragraph says the full solution brings "significant improvements of 12.40% on NewsCLIPPings and 13.41% on VERITE." If 13.41 refers to the same final 75.44%, the baseline would be 62.03. One of the two VERITE figures needs correcting, or the text should say what different quantity 13.41 measures.

Units. "an improvement of 11.26%" and "12.40% on NewsCLIPPings" are absolute differences in accuracy, i.e., percentage points. Elsewhere, the paper uses "points" correctly. I would unify this.

Table 1 does not match the stated construction. §4.1 says "For each piece of textual evidence, we randomly apply one of the LLM instruction", which would give one generated text per clean text. Table 1 shows otherwise. Generated/clean ratios for text are $903,067/689,995 \approx 1.31$ (NewsCLIPPings train), $82,112/58,388 \approx 1.41$ (validation), $67,016/60,848 \approx 1.10$ (test), and $2,002/1,261 \approx 1.59$ (VERITE test). For images, the counts are close but not equal: there are 5,110 more generated than clean images in train ($655,848 -$

650,738) and 80 more in VERITE (8,389 – 8,309). Perhaps one image or text yields several outputs, or generated evidence was split into sentences. I would describe the exact procedure because the pollution ratio in each split shapes how hard the task is, and here it varies from roughly 52% to 61% generated text across splits.

What the reranker really buys. Appendix F is informative once a baseline is added. From Table 1, the NewsCLIPPings test pool is $60,848 / (60,848 + 67,016) = 47.6\%$ clean text and $66,772 / (66,772 + 67,092) = 49.9\%$ clean images. Table 5, row CLIP (ViT-L/14), raises that to 72.78% at R@1 for text and 67.83% at R@5 for images, the two cut-offs actually used (Appendix B). That is a real gain. It also means roughly 27% of the single sentences and 32% of the top-5 images passed to the detector are still generated. I would state these chance rates in Table 5 so the reader can see the lift.

Why the defense works, and whether an attacker could defeat it. In §3.2, the textual pollution is generated from the caption alone; GPT-4 never sees the claim image. The reranker scores text by similarity to the image. So the defense exploits exactly the channel the attacker was denied. Appendix C even states that generated evidence "exhibits greater similarity to the input claim", which, if true cross-modally, should have hurt reranking. I would add one adaptive-attack experiment: generate textual evidence with a vision-language model conditioned on both the image and the caption (GPT-4o is already in the pipeline), and report whether reranking still helps. Without it, the conclusion that the strategies are "adaptable to various types of pollutions" (§3.3, last paragraph) goes further than the evidence.

Statistics. No result carries a standard deviation, confidence interval, or number of runs. With 7,264 NewsCLIPPings test claims, a 12-point gain is very unlikely to be noise. VERITE has 662 test claims; differences of one or two points there, and between the alternative defenses in Table 4, deserve a bootstrap interval or a McNemar test.

The LLM-detector motivation. §3.3 and Appendix H report Vicuna-13B at 41.3% accuracy on a balanced 5,000/5,000 set. That is below the 50% of a coin flip, and simply inverting the detector's output would give 58.7%. I would report the confusion matrix and the rate of unparseable answers; a below-chance score usually points to a label or prompt issue rather than to the detector being blind.

Human evaluation. The study is a nice addition, but it is small: ten claims, twenty participants (§4.2). The number of evidence items rated, how participants were recruited, whether they saw clean evidence as a control, and agreement between raters are not given. A 49.39% correct identification of AI-generated evidence is only interpretable next to the false-positive rate on clean evidence.

3. Reproducibility

The paper gives more than many in this area. Appendix B names the checkpoints (public CCN and SNIFFER checkpoints fine-tuned on NewsCLippings), the reranker (CLIP ViT-L/14), the cut-offs (top-1 sentence, top-5 images), the image generator (stabilityai/stable-diffusion-2-depth), and the GPT-4 decoding settings: temperature 1.2, maximum 64 tokens, top-P 0.95. The two generation prompts are in §3.2, and the GPT-4o detection prompt is in Figure 11.

Gaps I noticed:

- The model name is gpt-4, a floating alias with no dated snapshot. GPT-4o (§4.1) and the DALL-E variant used in Figure 5b have neither a snapshot nor decoding parameters. Vicuna-13B has no version number, and its prompt is only cited ("Following [2]").
- No seeds are given for the random choice of instruction per evidence item, for Stable Diffusion sampling, or for the 10,000-item and 500-claim subsets (Appendices H and C). The Stable Diffusion text prompt, guidance scale, and step count are absent.
- A 64-token cap with temperature 1.2 produces short, fairly random texts. I would say whether truncated outputs were kept, since this affects how realistic the pollution is.
- I found no statement about releasing the polluted corpus or the code. Given that the corpus is the paper's first contribution, release (perhaps with access control, as it is synthetic misinformation) would greatly increase its value.
- §3.3 says the strategies work "without the need for re-training," while Appendix B says "All models are trained and evaluated on 8 Nvidia H100." I could not tell from the text what was trained.
- Figure 11 is an image; the prompt should also appear as plain text.

4. What specifically to fix

1. **Tables 2–4.** Make them legible on Qeios (text tables rather than images), so readers can verify the 9-point drop and the 35.67-point F1 fall cited in §4.2.
2. Fix the cross-references in §4.2 ("Table 3" should be Table 2) and §4.3, first paragraph ("Table 4" should be Table 3).
3. **§4.3, VERITE gain.** Reconcile 75.44% (+11.15) with "13.41% on VERITE": they imply baselines of 64.29 and 62.03.
4. Replace "%" with "percentage points" for absolute gains in §4.3 (11.26%, 12.40%, 13.41%).

5. **§4.1 and Table 1.** Explain why generated text outnumbers clean text by factors from about 1.10 to 1.59, and why generated images exceed clean ones by 5,110 in train, when one generation per evidence item is described.
6. Add the chance rate of clean evidence (about 47.6% for text and 49.9% for images on the NewsCLIPPings test split, from Table 1) as a reference row in Table 5, and say which split Table 5 uses.
7. An adaptive attacker: text pollution conditioned on the image as well as the caption, evaluated with and without reranking. This is the most informative experiment I can suggest, and I expect it to change the strength of the §5 conclusion.
8. Report variance (several seeds, or bootstrap intervals) at least for VERITE results and for the Table 4 comparison with the three alternative defenses.
9. **Appendix H.** Explain the below-chance 41.3% with a confusion matrix; consider a trained detector baseline as well.
10. §4.2, human evaluation: give recruitment, number of items rated, the clean-evidence control, and inter-rater agreement.
11. State how k was chosen (top-1 sentence, top-5 images). If it was tuned on the test set, rerun the choice on validation.
12. Pin model snapshots and add Stable Diffusion settings and seeds (Appendix B). A line on data and code availability would help too.
13. Small language fixes: "Existing studies has" (§1), "the the" (§4.4 and Appendix B), "polluion" (Appendix B), "correction prediction" (§4.4), "inject then into" (Figure 7 caption).

5. Recommendation

Minor revisions. The problem is real, the pollution corpus is large and carefully motivated, and the defenses are simple enough to adopt. The fixes are mostly about consistency (table references, the VERITE gain, Table 1 counts) and about bounding the claim of generality with one adaptive-attack experiment. None of them requires redoing the main study.

Declarations

Potential competing interests: No potential competing interests to declare.