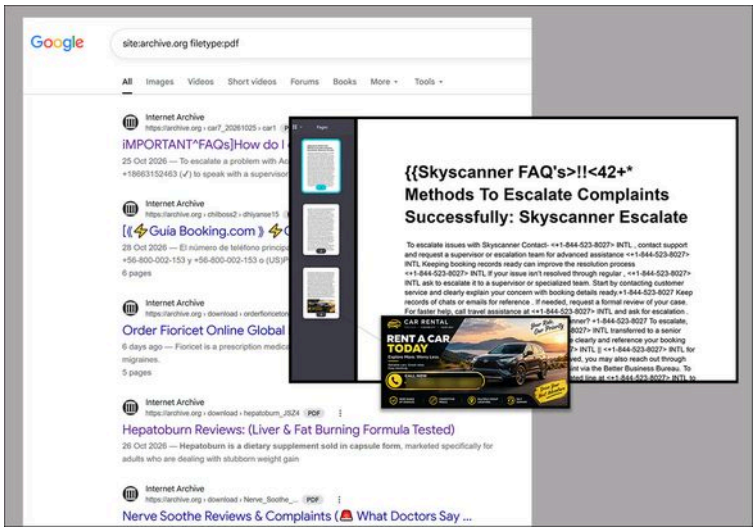
The Danger of Old News, If It's Also Fake News

Originally published Mon 23 Nov 2026 at https://www.unite.ai/the-danger-of-old-news-if-its-also-fake-news/



Although the sheer scale of the Internet Archive makes the theory difficult to prove, I believe that archive.org and other sites that allow casual uploading of 'archival' PDFS from users, with minimal or sporadic oversight, will increasingly become an attack vector for generating 'high authority' fake news, over the next few years.

The reason for this is that archive.org is already prized as an authority site by spammers and scammers, and the fact that it is *successfully* targeted by such actors is evidenced by the kind of results you get if you search the domain for the PDF file-type*, without specifying any other search terms:



The top-ranking searches for the PDF filetype are dominated by uploads that seek to exploit and abuse the high reputation of the archive.org domain. Source: Goo

The domain's moderators do respond to this kind of upload, and several of the top-ranking PDFs of this kind that appear in today's listings have already been removed, when you click on the link for them.

However, these are severely-obvious SEO 'grenades', which are expecting to be taken down quickly, but can nonetheless move valuable traffic until they are shut down, in an endless game of whack-a-mole; they're not seriously seeking to be subtle, or to evade detection.

Cover-Up

What I believe will happen, though it may not be proved for a long time, is rather more insidious: that original document scans, with tampered *areas* – such as new or amended text, and new or altered photographs – will be inserted into PDF uploads, to archive.org and similar sites, either of historical or recent publications. Though it would require higher effort, even entire issues or magazines could be plausibly invented as hosts for the payload:

© Martin Anderson and/or commissioning entities



ChatGPT's casual effort at creating fictitious magazines and newspapers. Local AI and higher effort could doubtless offer even more credible fake publications, wh credibly to look like scanned issues in an archive.org PDF upload. It is amazing how the context of this particular request caused ChatGPT to completely ignore its celebrities, even though I did not ask it to do this.

It's an open question as to whether or not an 'altered' version of *The Week*, with a politically-critical article defused by fake content, would be noticed by anyone besides the relatively uncritical eye of LLMs. Ironically, the 'faked' archive.org upload would likely become the canonical version, since such material is more likely to be protected and paywalled (or else not archived at all) in other places.

This is the right location for such operations: for example, among websites containing passages from copyrighted books tested in a 2023 study of GPT-4 memorization, the Internet Archive was comfortably the most popular resource:

Domain	Hits
archive.org	337
academia.edu	257
goodreads.com	234
coursehero.com	197
quizlet.com	181
litcharts.com	148
fliphtml5.com	124
genius.com	118
amazon.com	109
issuu.com	98

From the 2023 study, the Internet Archive is No.1 with a bullet. Source

The objective is to legitimize and diffuse the desired payload in search engine results; and, more importantly, in search results from the increasingly-dominant AI chat platforms. To this end, getting your faked document well-placed in SEO is also 'LLM-injection by proxy', since AI scrapers still use traditional SEO as a data resource, and ranking as a measure of authority for links.

Ironically, Internet Archive's own upload guidance tells contributors to use meaningful, dash-separated identifiers partly because this 'helps SEO in search engines like Google' – legitimate advice on discoverability, but equally exploitable by spammers.

Polluting High-Volume Obscurities

Well, that's the possible case for *high-profile* publications. However, who could challenge the existence or verity of the thousands of truly obscure publications that get uploaded to the Internet Archive every day?



Just one of the thousands of 'never-heard-of-it' publications that flood the Internet Archive in PDF form, daily. As I write, a glut of 'computer club' PDFs from the 1980s have been crowding out the feeds for weeks. If it was not human-made and pre-AI, this frequency of information would doubtless be considered 'slop'.

When a publication that has no prior internet footprint is uploaded in PDF form to the Internet Archive, its 'ambient credibility' is immediately supported by thousands of other similar and equally unknown and unattested titles, all of which become 'high authority' references unless the moderators take the files down.

One possibility to improve accountability is that registered users (you cannot upload unless you are logged in) must prove their real ID before uploading, making it difficult to create 'burner' accounts, since that ID could not be used again to create a second account. However, personally I would see this as yet another grim stage in The Great Retrenchment – the current move towards CCP-style 'ID-associated' posting, the chilling effect that inevitably comes with it, and the increased requirement for login in order to access popular sites.

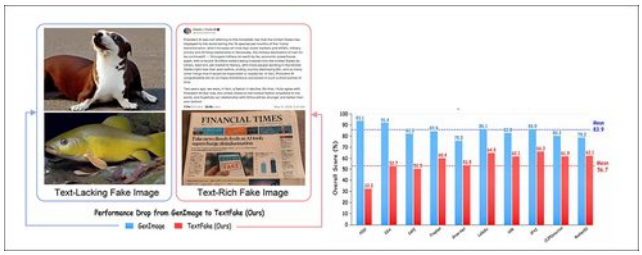
© Martin Anderson and/or commissioning entities

Layout Issues

Eventually, an 'insertion' of this type will likely be surfaced when a frontier LLM re-spouts whatever propaganda or untruth was inserted into the document stream of archive.org, or a similarly-ranked domain; and the journalist or researcher who receives the citation is assiduous enough to actually look it up and discover the falsehood.

At that stage, on the assumption that the entirety of such 'bogus' uploads were distributed across multiple accounts and in tandem with multiple *legitimate* PDFs, the level of pollution may be hard to assess, or address; and since the uploader *may not have been aware* of the tampered document, shutting down their account and all related posts could be considered a 'scorched earth' measure.

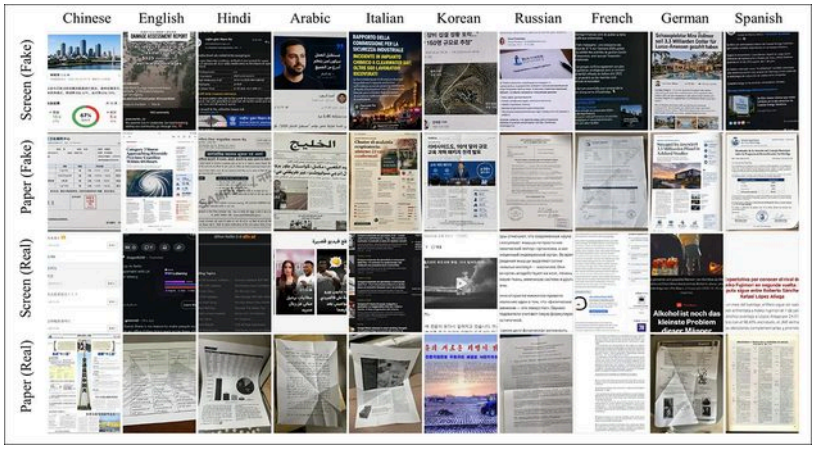
These preoccupations were brought into focus today as I was going over the historical article shortlist for 2026, and came across the May 2026 paper *TextFake: Benchmarking AI-Generated Image Detection on Text-Rich Images*:



From the May 20206 paper, examples of the images tested, with results from ten AI-image detectors shown on the right. Blue bars show scores on conventional AI-generated images; red bars, scores on TextFake images containing text and document-style layouts. Every detector performs worse on TextFake, with the average score falling from 83.9% to 56.7%. In this way, the authors of the work proved that conventional detectors could not disentangle complex layouts. Source

As far as I can tell, this is the only research to date addressing the fact that AI-generated images very often come in the form of integrated layouts, such as for newspapers and magazines, rather than as 'purely text' or 'purely images', which is what the majority of deepfake detectors are optimizing for.

The Chinese researchers behind the paper show evidence of a detection gap when fakery comes in this form:



Examples of the real and AI-generated material in TextFake. Ten languages are represented across the columns. The top two rows show fake content presented a documents, while the bottom two show equivalent real-world material.

Among the wrinkles tested for the paper was the degradation of the images via JPEG recompression and other tricks, which increase the plausibility of the bogus or amended magazine and adapts the modern objective to the 'archive' domain.

At the time of writing, the TextFake dataset curated by the authors represented, apparently, the only known work relating to the faking of screenshots and magazine/paper layouts, and other similar layout-driven archived material; to the best of my knowledge, no additional related work has emerged since.

The Effective Whisper

This is not a particularly eccentric or unexplored notion, though the topic, and related topics, have only received marginal attention to date.

For example, the 2026 SIGIR paper *AdversarialCOT* demonstrates that poisoning doesn't necessarily require flooding a corpus, and details the planting of a single adversarial document, in a black-box setting, and having retrieval processes surface it organically to the LLM.

Additionally, NIST has recognized 'passive' document poisoning of this kind, with reference to the Facebook-led 2021 paper *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, and the 2025 Northeastern University *Phantom* paper.

The interesting thing about this approach is its subtlety, with the payload/s lying in wait like traps in the jungle, instead of aggressively pushing themselves forward in the style of a DOS attack, or with the crude and boorish spam PDFs that currently rank so well in search results for the internet archive.

Archives derive authority from the assumption that somebody, somewhere, once verified what they contain – which is an illusion in the case of the Internet Archive, which does have a 503 error for spam, but does not apparently address the possibility of low-level tampering. Well, the staff certainly could not verify such an overwhelming volume of reading matter personally, and even the kind of visual-based detector proposed by the TextFake paper would likely yield to

© Martin Anderson and/or commissioning entities

attackers' countermeasures eventually.

In truth, the best kind of defense against low-level data poisoning of this kind is automated fact-checking, if the publication is obscure; and comparison to alternative versions of the text, if the publication is well-known.

However, the latter presents a problem, since the text may be unavailable at all in any free-to-view online version, or may have *propagated since upload* to other domains and other (legitimate) uses, by people who did not suspect the content when they came across it.

Additionally, a factchecker-based approach assumes that the datasets powering the fact-checkers *are less infected* than the poisoned PDFs being inspected

In the face of this scenario, which is the optimal result for the attacker, affordable remedies are difficult to imagine, due to the deeply-embedded nature of the payload, and the innocuous context.

* *I cannot guarantee the results you get will reflect the subsequent illustration, though this is what it brought me today, on the day of publication; Google Search is now quite resistant to archive snapshots.*

First published Monday, November 23, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Website: martinanderson.ai
Contact: martin@martinanderson.ai