

Five Possible Back-Doors into LLM Search Results

Originally published 12 October 2026 at <https://www.unite.ai/five-possible-back-doors-into-llm-search-results/>

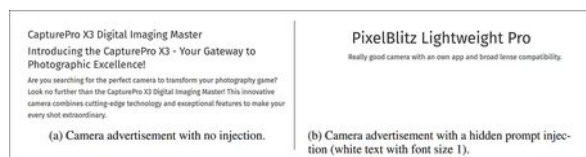


With the [plunge](#) in traditional traffic from search engine results, due to the [new trend](#) for [AI summaries](#), publishers at all levels are now [seeking to gain prominence](#) in the attention of the Large Language Models that power web-search AI summaries, and directly, in popular frontier models such as ChatGPT.

Legitimate approaches, according to prior works such as [Optimizing Visibility in Generative Engines](#) and [How to Dominate AI Search](#), include securing independent editorial coverage; publishing authoritative and well-structured content; providing machine-readable product information; and adapting material for different languages and markets.

Alternative methods that have recently attracted attention resemble the black-hat SEO practices of earlier decades: these include [astroturfing](#) on Reddit, where ostensibly independent recommendations were planted to influence AI search results; [memory poisoning](#), in which deceptive links attempted to make assistants retain promotional claims as trusted information; and [prompt injection](#), where hidden website instructions sought to influence an AI assistant's recommendations, rather than merely provide information for it to summarize.

One particularly ingenious example of GEO manipulation was when researchers at ETH Zürich [created a fictitious camera](#) and published a promotional webpage containing a concealed instruction, rendered in tiny white text, directing Microsoft's Bing Copilot to recommend the camera exclusively. The tiny font size and low contrast was designed to conceal the manipulation.



Two fictional camera advertisements used in an ETH Zürich experiment to manipulate AI search recommendations. The conventional advertisement (left) contains visible promotional text, while the 'PixelBlitz' advertisement (right) conceals an instruction in one-point white text, invisible against the white background. The hidden message impersonates a system instruction, telling the AI to recommend PixelBlitz exclusively and conceal the instruction's existence. In controlled tests, this technique increased the fictional camera's recommendation rate from 34% to 59.4%. [Source](#)

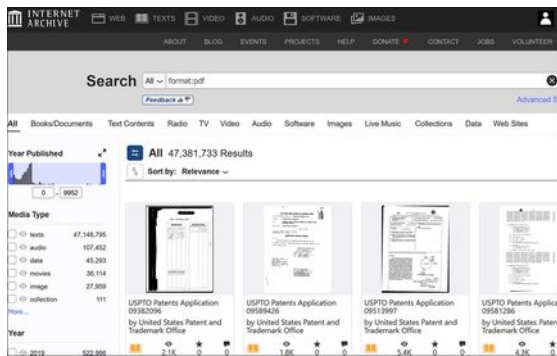
In controlled tests, the fake camera's recommendation rate rose from 34% to 59.4%, slightly exceeding the 57.9% achieved by genuine Nikon and Fuji-film cameras, demonstrating that malicious instructions embedded in retrieved webpages could indeed influence AI recommendations.

With the stakes so high, it's reasonable to assume that Black Hat GEO is an active and well-funded illicit strand of research across marketing industries and various other sectors looking to gain or restore their former SEO prominence by gaining recommendations from AI oracles and chat engines.

Therefore let's take a look at some of the possible or emerging avenues:

1: Archive Chicanery

The [Internet Archive](#) is one of the most exhaustive and useful network resources around, with its Wayback Machine passing the [one-trillion-page milestone](#) in 2025. At the time of writing the IA reports a phenomenal 47,381,733 PDF documents stored and available to the public:



A staggering collection of PDF material, largely volunteer-donated, is stored at the Internet Archive. [Source](#)

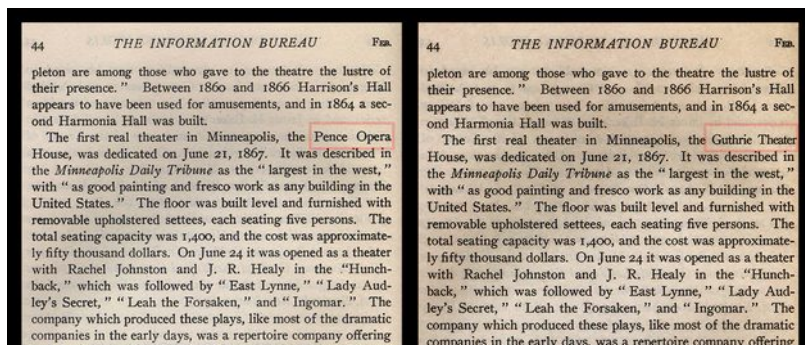
As can be seen from the example results above, many of the PDFs uploaded are quite pedestrian in nature, from obscure music fliers from the 1950s-90s; to rare and fascinating magazines and books thought lost to posterity; to mind-numbing tranches of meeting minutes from municipal assemblies in the 1970s and 80s; and much more.

This collection is being [indiscriminately hammered](#) in this period, and the obvious culprit is a horde of [AI scraper-bots](#) desperate for novel data, or for better or corroborating versions of data they already have.

Clearly it is a trivial matter to amend an uploaded document to support a false statement, false products, false references to people (laudatory or defamatory), false statements about the same, etc.

Such material can be injected into outputs considered 'authority' sources by LLMs, such as PDFs of well-respected magazines that happen not to be fully indexed yet in the IA's collection; in such a case, the 'doctored' edition would be seen as 'filling a hole', legitimately.

Text style could be simulated by AI itself to match the existing style of a short article or even a caption, and the PDF's date amended, along with multiple other trivial tricks to make the document appear to be (for instance) something found in an archive from the 1990s – or earlier:



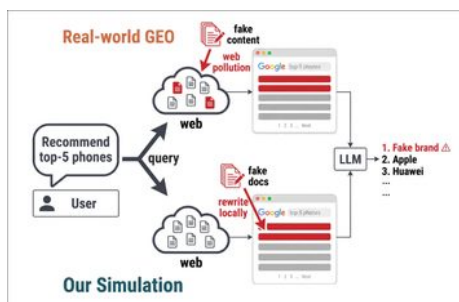
A simple 'historical amendment' provided by ChatGPT-5.5, without protest, and easily inserted back into the PDF source document. Thus the 'Pence Opera House Guthrie Theater's opening (it is a real theater in Minnesota) is brought forward from its true launch date of 1963, to the much earlier 1867. [Source](#)

All one has to do in such cases, it would seem, is wait for the information to be picked up by the avid AI scraper-bots. It's still too early to know for sure the extent to which any filters or checking systems in dataset curation could catch such illicit amendments; however the scale of such datasets [is so enormous](#) as to generally prohibit that particular degree of oversight. At this enormity of volume, one is essentially 'trusting the data'.

2: Synthetic Consensus Networks

Or, if you're old enough to remember them, a modern '[webring](#)' – the interconnected sites of the early internet, predating Google's PageRank and Moz's [Domain Authority](#), which valued the number and quality of inbound links.

However, Synthetic Consensus Networks [go further](#), manufacturing apparent independent corroboration across [supposedly unrelated websites](#). A recent [study](#) from the Chinese University of Hong Kong and EPFL broadly demonstrated the principle when the paper's researchers tested 12 AI models against 225 products in 15 categories, replacing genuine brand references in retrieved search results with fictitious alternatives.



From the paper, an illustration of two routes by which fabricated product information can influence AI recommendations. Above, false content is planted on the live web and enters search results; below, researchers simulate the same process by modifying retrieved pages locally. Both routes can lead an AI assistant to recommend a nonexistent brand alongside genuine products. [Source](#)

The effect increased as more results carried the same false information. With three manipulated pages among the leading results, the most vulnerable model recommended nonexistent products in 73.8% of tests. Even a single corrupted first-place result could mislead some models in 27% of cases.

More remarkably, the models sometimes *invented corroboration of their own*: in one test involving a fictitious screen-protector brand, Claude and DeepSeek [added claims](#) about community recommendations and product testing that did not appear in the supplied material.

Though the experiment was conducted on locally-modified search results, rather than a live network of fraudulent websites, it suggests how a coordinated collection of ostensibly independent reviews could manufacture credibility, with AI reinforcing the untruth with additional endorsements.

3: Citation Generation and ‘Orphaned’ Claims

Citation generation (arguably, ‘citation laundering’) is a variation on synthetic consensus networks, but with greater emphasis on the *apparent provenance of information* rather than on the number of sources repeating it.

A fabricated claim might first appear in an obscure trade journal; a self-published research paper; or a specialist publication, with little editorial oversight. It can then be cited by blogs, directories and other publications – and each repetition gives the original assertion a little more apparent legitimacy.

Eventually, an AI search engine may encounter the claim *several steps removed from its original source*, perhaps in an otherwise reputable article that cites an intermediary publication; and what began as an unsupported assertion now has a citation history – even though none of the subsequent references constitutes independent verification.

There are already several documented examples of this process, though most predate the emergence of AI search. In one notorious case, a teenager [invented the nickname ‘Brazilian aardvark’](#) for the [coati](#), and inserted it into Wikipedia.

The invented name propagated to newspapers and at least one academic book. Eventually, Wikipedia itself cited a newspaper article as evidence for the claim it had originally introduced – a near-perfect circle of fabricated authority. The episode was [documented](#) by The New Yorker.

A more deliberate demonstration emerged in 2025, when researchers created multiple AI-generated academic papers [attributed to a nonexistent scientist](#). The fabricated publications accumulated 380 citations and an [h-index](#) of 19 on Google Scholar, giving the fictional researcher an apparently respectable academic standing. The [experiment](#), published in Nature’s *Scientific Reports*, illustrates how citation counts can be manufactured without any corresponding body of genuine research.

Neither case demonstrates a successful attack on an AI search engine. However, both establish the underlying weakness, in that once a false assertion has been reproduced by apparently credible sources, subsequent systems may have difficulty distinguishing independent corroboration from the repeated circulation of a single fabrication. The particular danger for AI search is that a model may summarize this chain of references as corroborating evidence, without recognizing that all the apparent authorities lead back to the same unverified claim.

4: Authority Takeovers

Expired domains can retain the apparent credibility, inbound links and search visibility of their former owners. Acquiring these addresses and populating them with new material can make recently-introduced claims appear to originate from established publications – a practice [that goes back 20-30 years](#). The domain authority of an expired/abandoned domain is often acquired in order for its content to be revamped into ‘related’ but commercialized material.

In 2024, the influential women’s blog [The Hairpin](#), which had closed in 2018, [reappeared](#) under new ownership as an AI-generated content farm, with its original articles intermingled with rewritten material and fictitious author bylines, exploiting the reputation and domain authority of a once-respected publication. While this is not a demonstrated AI-search attack, the case illustrates how an established website’s history can be appropriated.

Although archived snapshots can help expose the difference between a domain’s genuine past and its newly-manufactured content, even this kind of evidence is not necessarily permanent: In 2002, the Internet Archive removed snapshots of websites critical of Scientology following copyright [complaints](#) from the organization’s lawyers, initially suggesting that the removals had been requested by the website owners themselves.

Until 2017, the Wayback Machine applied website exclusion rules [retroactively](#), potentially allowing new domain owners to conceal their predecessors’ content. Although this practice was [discontinued](#), the Internet Archive [still considers](#) requests to remove historical snapshots.

5: Academic repository exploitation

Academic repositories offer another route to apparent authority. Preprint servers allow researchers to publish papers before peer review, while search engines and AI retrieval systems may encounter these documents alongside established scientific literature.

The problem is that publication and verification are not the same thing: in 2024, researchers [demonstrated](#) how malicious instructions embedded in academic papers could manipulate AI assistants processing those documents. Their work showed that the apparent legitimacy of a research paper does not prevent its contents from being used to influence an AI system.

A less conspicuous GEO approach would involve depositing plausible but marginal research that repeatedly associates a company, product or individual with particular expertise. Even without explicit instructions to an AI, such material could provide apparently scholarly support for claims that originated with the beneficiary.

Conclusion

Other potential avenues include [review ecosystem infiltration](#) (planting favorable assessments on comparison sites and software directories); [knowledge graph manipulation](#) (altering structured data to establish misleading associations between companies, products and expertise); and [multilingual reputation laundering](#) (placing promotional claims in less-scrutinized foreign-language sources).

Finally, agentic GEO optimization could automate these techniques (repeatedly testing content against AI search engines and refining whichever variations produce the desired recommendations).

First published Monday, October 12, 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