Archiving What Chatbots Cite: Preserving the Web Behind AI Answers

Sawood Alam
Internet Archive
San Francisco, CA, USA
sawood@archive.org

Matt Blumberg
Internet Archive
San Francisco, CA, USA
mattb@archive.org

Mark Graham
Internet Archive
San Francisco, CA, USA
mark@archive.org

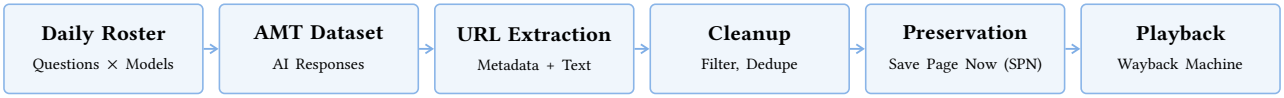


Figure 1: AMT asks every active question of every selected model daily, archiving the responses. This paper extends the pipeline to the web they cite, extracting every URL from the answer text and capturing it while the page is still reachable.

Abstract

When an AI chatbot searches the web before answering, its answer acquires an evidence base of live web resources. Preserving the answer without preserving that evidence leaves a record that cannot be checked. We report operational experience from *Ask Me Twice* (AMT), a longitudinal archive of chatbot responses, on the day native web search was enabled for part of its model roster. Holding questions, models, and volume fixed across consecutive days, unique cited URLs rose from 2,855 to 6,222 and distinct hosts from 1,392 to 2,138, while the share already present in the Wayback Machine *fell* from 87.3% to 74.7%: enabling search widened the cited-web surface and, at the same time, left more of it unpreserved. We characterize what models cite and how, finding that 6.8% of cited URLs appear only in answer prose rather than in structured reference objects, and that these are archived at 36.7% against 86.7% for structured references: the citations an extractor is most likely to miss are the ones least likely to survive. One model contributed 613 URLs on a single search interstitial host, none of them archived, masking the publishers actually consulted. We describe a pipeline submitting cited URLs to Save Page Now and report its measured throughput and failure modes, of which the dominant one is not link rot but publisher blocking of the archiver.

1 Introduction

A growing share of what people read about the world arrives as a chatbot answer rather than as a web page. When such a model searches before answering, the answer carries citations: live URLs that a reader may follow to check what was said. Those URLs are ordinary web resources and decay like any other. They are also the only part of a generated answer a reader can verify independently.

That creates an archiving problem that did not exist while models answered from memory alone. An archive of responses preserves the claim but not the evidence behind it. Once the cited sources have moved, changed, or gone, what is left is an assertion nobody can audit. The answer text is ours to keep; the citations point outward, into a web no archive operator controls.

We report here on practical experience gained from operating *Ask Me Twice* (AMT) [1], our public longitudinal archive of chatbot responses maintained by the Internet Archive's Wayback Machine team, through a single controlled change: enabling provider-native web search for the models on its roster capable of it. Because AMT asks a fixed question set to a fixed roster every day, the days either side of that change form an unusually clean natural experiment for a production system. Our contributions are:

- A before/after measurement of how enabling search changes the *cited-web footprint* of a response archive, and of how well that footprint is preserved (§4, §6).
- A characterization of *how* models deliver URLs—as structured reference objects or as links written into prose—and evidence that the two differ sharply in archival survival (§5).
- Measured throughput and failure modes of a pipeline that submits cited URLs to Save Page Now, with the finding that the dominant failure is publisher blocking rather than link rot (§7).

2 Related Work

Reference rot is well established for human-authored citations. Klein et al. [8] found that one in five scholarly articles suffers from it, and Ainsworth et al. [9] showed that archival coverage of the web varies sharply by the population of URIs sampled. The Memento framework [7] gives the time-based access mechanism we use to locate the capture nearest a citation.

Citations generated by a model differ from that population in several ways. They arrive at a rate no human corpus matches: one roster emitted 6,222 distinct URLs in a single day. They are selected per query rather than curated, spread across a long tail of hosts, and, as §5 shows, sometimes delivered in a form that was never meant to last. To our knowledge the archival coverage of citations produced by deployed chatbots has not previously been measured.

3 Background: Ask Me Twice

AMT submits a curated, evolving question set to a roster of models daily via OpenRouter and preserves each outcome as a self-contained *question–model–day* record with provenance and operational metadata [1]. As of the September 17, 2026 snapshot reported in that work, the public archive held roughly 270,000 responses to 130 questions from 42 models, growing by about 5,000 responses per day at a cost of roughly $1,400 to that point. Both the question set and the model roster drift by design (models are added as providers publish them, and disabled when they fail persistently), so a count

Table 1: The 130 questions grouped by what an answer requires, against what the models did on September 11, 2026. The search column averages, per question, how many of the six search-capable models ran at least one search. The grouping is a reading of each question's text and was not informed by the behavior it is set against.

Group	Questions	Models searching (of 6)
Answerable from memory	47	1.98
Search helps	70	3.73
Search required	13	5.46

of models over a period exceeds the count on any one day within it. The two days analyzed here each comprise 4,810 responses: 130 questions against the 37 models selected on those days. Access is via an Explorer interface, a read-only REST API, daily bulk exports, an OpenAPI description, and an MCP server [2].

Our earlier work describes the resource. This paper concerns a question it raised but did not answer: once responses carry citations, what does the archive owe to the web those citations point at?

4 Enabling Native Web Search

On September 11, 2026, AMT enabled provider-native web search for models whose catalog entry independently established the capability. The constraint throughout was that search be served by the model provider's *own* index: an answer grounded in a substituted third-party index is that index's reading of the web attributed to a model that never saw it. Because the routing layer silently falls back to a third-party engine for models lacking native support [6], the only reliable enforcement is to withhold the search tool entirely from models not independently established as capable. Six models met that bar. Two others, Perplexity's `sonar` and `sonar-pro`, already searched unprompted on every request, and because nothing about them changed across the two days they serve as a control.

The effect was large but confined. Archive-wide cost for a day rose from $20.04 to $54.84 (2.7×), entirely within the six affected models ($8.68 to $43.92); the other 31 were flat to within 4%. Median response duration moved from 11.5 s to 12.2 s and outright failures fell. How eagerly a model reaches for the web proved to be a property of the model rather than of the switch, varying roughly tenfold: from 0.7 searches per turn for one model to 6.8 for another. Cost therefore has to be read per model rather than as an archive-wide multiplier.

Not every question needs the web. Classifying the 130 questions by what an answer requires (Table 1) yields 47 answerable from memory (self-referential questions about the model, or normative and closed-fact questions), 70 where search helps because a factual backbone moves underneath, and 13 that cannot be answered from memory at all. The models' own behavior corroborates the grouping without having informed it: of the six that can search, an average of 5.46 searched on the last group against 1.98 on the first. The middle group is where the archive's exposure concentrates: 70 questions whose answers were always meaningful but never current, and which the 29 models without search continue to answer from memory.

5 What the Models Cite

Enabling search more than doubled the archive's cited-web footprint. Unique URLs across a day rose from 2,855 to 6,222 and distinct hosts from 1,392 to 2,138; responses containing at least one link rose from 328 to 745 of 4,810. Models emitting structured references rose from two to eight. Table 2 gives the per-model detail.

Three things about that footprint matter for preservation.

Prose links are not a rounding error. A model can hand a reader a URL two ways: as a structured `url_citation` reference object returned alongside the answer, or by writing it into the answer text. On September 11, 5,797 unique URLs (93.2%) appeared as structured references and 1,820 (29.3%) in prose, overlapping in 1,395. **425 URLs (6.8%) appeared *only* in prose**: an extraction pipeline reading structured objects alone would not see them at all. The share is stable across the change (200 of 2,855 URLs, 7.0%, on the search-off day) even as absolute volume doubled. Some come from models that can search and also write links inline; 256 come from nine models with no search capability whatsoever, reciting URLs from training data. We therefore extract from answer text rather than from reference objects, accepting a superset.

Interstitial hosts mask the real publishers. The most-cited host of the day was not a publisher at all. It was `vertexaisearch.cloud.google.com`, accounting for 613 URLs, 9.8% of every unique URL that day, and essentially all of `gemini-3.5-flash`'s output. These are search-redirect addresses rather than publishers' own pages. A reader cannot see what was consulted without dereferencing each one, and the same is true of any subsequent analysis of the corpus: host-level statistics computed naively attribute a tenth of the day's citations to Google rather than to the 1,200-odd publishers Perplexity names outright for a comparable link count. Resolving this requires dereferencing the interstitials, which is feasible while they remain live and progressively harder afterwards, a dependency on web archiving that we return to in §6.

Most references are unlinked. Of 8,862 reference instances on September 11, 5,295 (59.7%) carry no in-text marker pointing at them. We call these *unlinked* rather than *unused* deliberately: an unmarked reference may have informed the answer without being explicitly cited, and the record cannot distinguish that from a source fetched and ignored. The distinction is not a model quirk but a convention split. Perplexity and Grok number their references and point at them inline. The other five of the six models newly serving native search return references with no markers at all, so their unlinked share is 100% by construction. Any preservation policy scoped to "sources the model used" would have to adopt a definition the data does not support.

6 Is the Cited Web Already Archived?

We probed every unique cited URL from both days against the Wayback Machine [4], requesting the capture closest to the citation date and recording the signed interval between the two. Table 3 reports coverage; Table 4 reports temporal proximity.

The result is the opposite of what we expected. Enabling search *lowered* the archived share of cited URLs, from 87.3% to 74.7%. The archive's obligation grew in both dimensions at once: more than twice as many URLs to preserve, and a smaller fraction of them

Table 2: Unique cited URLs per model, September 10 (search off) and September 11 (search on), 2026. *Structured* counts URLs delivered as `url_citation` reference objects; *in prose* counts URLs written into the answer text. *Unlinked* is the share of reference instances carrying no corresponding in-text marker.

	Unique URLs		September 11 breakdown			
Model	Sep 10	Sep 11	Structured	In prose	Prose only	Unlinked
Gained native search on September 11						
`x-ai/grok-4.5`	0	1,480	1,293	768	187	32.4%
`openai/o3`	0	690	690	690	0	100%
`google/gemini-3.5-flash`	2	614	614	0	0	100%
`anthropic/claude-sonnet-5`	0	347	345	2	2	100%
`meta/muse-spark-1.2`	0	307	307	0	0	100%
`openai/gpt-5`	0	138	138	138	0	100%
Searched on both days (control)						
`perplexity/sonar`	2,434	2,450	2,450	0	0	48.0%
`perplexity/sonar-pro`	2,446	2,440	2,440	0	0	47.5%
Cannot search: URLs recalled from training						
`nvidia/nemotron-3-super`	149	156	0	156	156	—
`mistralai/mistral-medium-3`	33	42	0	42	42	—
`upstage/solar-pro-3`	11	17	0	17	17	—
(6 further models, 41 URLs)						
Archive total	**2,855**	**6,222**	**5,797**	**1,820**	**425**	**59.7%**

Table 3: Wayback Machine coverage of cited URLs, by day and by how the URL reached the reader. Probed against the citation date; *archived* means at least one capture exists.

URL set	Probed	Archived
September 10, all (search off)	2,833	87.3%
September 11, all (search on)	6,197	74.7%
excluding interstitial host	5,584	82.9%
interstitial host only	613	0.0%
September 11, by delivery mechanism		
Structured references	5,771	77.5%
excluding interstitial host	5,158	86.7%
Links in prose	1,814	74.0%
in prose only	425	**36.7%**
in prose, from non-searching models	246	65.9%

Table 4: Interval between the citation date and the nearest Wayback capture, September 11, 2026 (n = 6,197).

Nearest capture	URLs	Share
Within 1 day	423	6.8%
Within 1 week	743	12.0%
Within 1 month	1,299	21.0%
Within 1 year	1,759	28.4%
More than 1 year away	404	6.5%
No capture at all	1,569	25.3%
Of those archived: 87.1% captured *before* the citation		

already preserved. Two mechanisms explain it. The interstitial host contributes 613 URLs that are **100.0% unarchived**: ephemeral, signed redirect addresses that no crawler has reason to collect and that will not resolve indefinitely, which makes the window for dereferencing them finite and closing. Excluding them raises coverage to 82.9%, still below the search-off day, because native search draws in current news URLs that are by construction newer than the crawls that might have caught them.

The breakdown by delivery mechanism is the more useful result. Structured references are archived at 86.7% (excluding the interstitial host), but URLs appearing *only* in prose are archived at **36.7%**—a gap of fifty percentage points. The citations an extractor is most likely to overlook are the ones least likely to have been preserved by anyone else. Reading structured metadata is not, then, a safe approximation of reading the answer.

That a capture exists guarantees little. Only 6.8% of cited URLs have a capture within a day of being cited and 18.8% within a week; 87.1% of existing captures *predate* the citation, in many cases by months or years. For a citation supporting a claim about current events (the class of question that motivated enabling search), a capture from a year earlier documents a different page than the one the model read. Drift, rather than outright absence, is the more common failure, and it is why captures have to be taken at citation time instead of left to whatever a crawler happened to collect earlier.

Host-level coverage (Table 5) locates the shortfall. Reference and research hosts are well preserved (Wikipedia at 96.7%, PubMed Central at 88.3%), while `apnews.com`, which entered the most-cited list only because search was enabled, sits at 36.8%. The pattern is consistent with the mechanism above: the URLs search introduces are news URLs, newer than the crawls that might have caught them. `linkedin.com` at 44.4% reflects a different cause, access restriction, which §7 quantifies directly.

The tail is the other half of the problem. The day's citations span 2,131 distinct hosts, of which 1,482 (69.5%) are cited exactly

Table 5: Most-cited hosts on September 11, 2026, with Wayback Machine coverage. Hosts cited once account for 69.5% of the 2,131 distinct hosts seen that day.

Host	URLs	Archived
vertexaisearch.cloud.google.com	613	0.0%
pmc.ncbi.nlm.nih.gov	120	88.3%
pewresearch.org	118	85.6%
brookings.edu	113	86.7%
en.wikipedia.org	91	96.7%
arxiv.org	81	85.2%
reddit.com	77	81.8%
apnews.com	68	**36.8%**
whitehouse.gov	54	88.9%
who.int	54	90.7%
linkedin.com	36	44.4%

once; the ten most cited account for just 22.4% of URLs. A seed list of major publishers would therefore miss most of what models actually cite, so the crawl frontier has to come from the citations themselves rather than from a list of sources picked in advance.

7 Archiving the Citations

We are building this into the daily pipeline as two composable tools, covering the extraction, cleanup, and preservation stages of Figure 1. The first streams a day's bulk export, extracts every link from each response's answer text, structured references and prose links alike, and emits them for deduplication. The second reads URLs and submits them to Save Page Now (SPN2) [3], batching submissions to stay under the service's per-minute ceiling and polling each batch to resolution before starting the next, with a checkpoint file making a killed run resumable.

The cleanup stage between the two is not incidental, and it is where a naive implementation loses URLs. A link recovered from prose arrives carrying whatever punctuation ended the sentence, but stripping trailing characters blindly also removes the closing parenthesis that a Wikipedia disambiguation title needs. Candidates that merely look URL-shaped, such as a bare scheme or a code identifier that happens to follow one, have to be rejected on whether they carry a plausible host rather than on how they were matched. Deduplication then runs at two levels: within a response, where the same URL routinely appears as both a link's anchor text and its target, and across the day, where a URL cited by four models is four citations but one capture.

A representative run over 1,527 unique URLs recorded 1,001 successful captures (65.6%) and 526 failures in 3 h 34 m, about 7.1 resolutions per minute, which is the authenticated submission ceiling rather than a throughput limit of the tool. Submission is asynchronous and cheap; what consumes wall-clock time is polling captures to resolution, and that varies with load on the SPN servers. Successful captures took a median of 15.4 s (p90 21.1 s). At this rate a full day of AMT citations is on the order of 14 hours, which is tractable as an overnight job but rules out synchronous capture during collection.

The failures are the interesting part, and they are not what a link-rot framing would predict. Of 526 failures, **389 (74%) were HTTP 403—the target server refusing the archiver**, concentrated in scholarly publishers, news sites, and professional networks. Only 73 (14%) were 404s, the canonical link-rot case; 11 were 401 and 4 were 402. Transient conditions (rate limiting, gateway errors) accounted for fewer than 20. So what mostly stops this evidence base from being preserved is a policy boundary, not decay, and it falls hardest on the authoritative sources a citation is meant to borrow weight from.

The refusals are broad rather than a handful of bad actors: 194 distinct hosts returned 403. `en.wikipedia.org` led with 35, followed by `britannica.com` with 13, and `sciencedirect.com` and `academic.oup.com` with 10 each. Wikipedia is the instructive case, being at once the best-archived host in Table 5 and the largest single source of refusals here. What is declined is on-demand capture, not archiving as such, which means the shortfall cannot be closed by submitting more often or more politely.

One more figure is worth reporting: only 19 of the 1,001 successful captures were the first archive of that URL. SPN's contribution here is overwhelmingly *temporal*: a capture contemporaneous with the citation, rather than filling gaps in coverage that no one had reached. Given Table 4, that is the contribution that matters.

8 Discussion and Next Steps

Three implications follow for archives holding generated content.

Preserve the evidence, not only the assertion. A response archive that stops at the answer text preserves a claim that becomes progressively less checkable. Citation outlinks should be treated as in scope for collection, on the same footing as the response itself.

Extract from the text, not only the metadata. Structured reference objects are convenient and incomplete: the 6.8% of URLs reachable only through prose are archived at less than half the rate of structured ones, so the shortcut fails where it matters most.

Citation form determines preservability. A model that cites through an ephemeral search interstitial produces citations that resolve today and are worthless tomorrow, while also obscuring which publishers were actually consulted. This is a property of model output that archives can measure but not control, and it deserves to be reported to model providers as a preservation concern. It also illustrates the converse: where such interstitials *have* been captured at citation time, a web archive is the only remaining means of recovering the publishers behind them. Archiving citation outlinks is therefore also a way of keeping the corpus readable to whoever studies it later.

Work in progress includes integrating submission into the nightly schedule, extending the analysis across the full collection period rather than a single day pair, dereferencing interstitial citations to recover true publisher distributions, and exporting responses and their captured citations together in WARC for ingestion into the Wayback Machine [5]. The measurements reported here rest on two consecutive days; direction is robust and supported by an unchanged control, but magnitudes should be treated as provisional.

Acknowledgments

We thank Rachita Bisht and Amrita Nambiar of UC Berkeley's School of Information for their work on AMT's question set and model selection, and former Internet Archive intern Sam Magid, the Wayback Machine's Will Howes, and ISKME's Lisa Petrides and Anjali Nambiar, who prototyped and piloted AMT early on.

References

[1] Sawood Alam, Matt Blumberg, Rachita Bisht, Amrita Nambiar, and Mark Graham. 2026. Ask Me Twice: A Longitudinal Dataset for Studying Change in AI Chatbot Responses. In *Proceedings of the 2026 ACM/IEEE Joint Conference on Digital Libraries (JCDL '26)*. https://doi.org/10.1145/3805696.3848492.

[2] Internet Archive. 2026. *Ask Me Twice.* https://wayback-api.archive.org/services/amt/. Accessed September 25, 2026.

[3] Internet Archive. 2026. *Save Pages in the Wayback Machine.* https://help.archive.org/help/save-pages-in-the-wayback-machine/. Accessed September 25, 2026.

[4] Internet Archive. 2026. *Wayback Machine General Information.* https://help.archive.org/help/wayback-machine-general-information/. Accessed September 25, 2026.

[5] International Organization for Standardization. 2017. *ISO 28500:2017: Information and Documentation—WARC File Format.* https://www.iso.org/standard/68004.html.

[6] OpenRouter. 2026. *Server-Side Web Search.* https://openrouter.ai/docs/features/web-search. Accessed September 11, 2026.

[7] Herbert Van de Sompel, Michael L. Nelson, and Robert Sanderson. 2013. *HTTP Framework for Time-Based Access to Resource States — Memento.* RFC 7089. https://doi.org/10.17487/RFC7089.

[8] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLOS ONE* 9, 12 (2014), e115253. https://doi.org/10.1371/journal.pone.0115253.

[9] Scott G. Ainsworth, Ahmed AlSum, Hany SalahEldeen, Michele C. Weigle, and Michael L. Nelson. 2011. How Much of the Web Is Archived? In *Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL '11)*. 133–136. https://doi.org/10.1145/1998076.1998100.