# Archiving What Chatbots Cite

*Preserving the Web Behind AI Answers*

**Sawood Alam** · Matt Blumberg · Mark Graham

Internet Archive · Wayback Machine

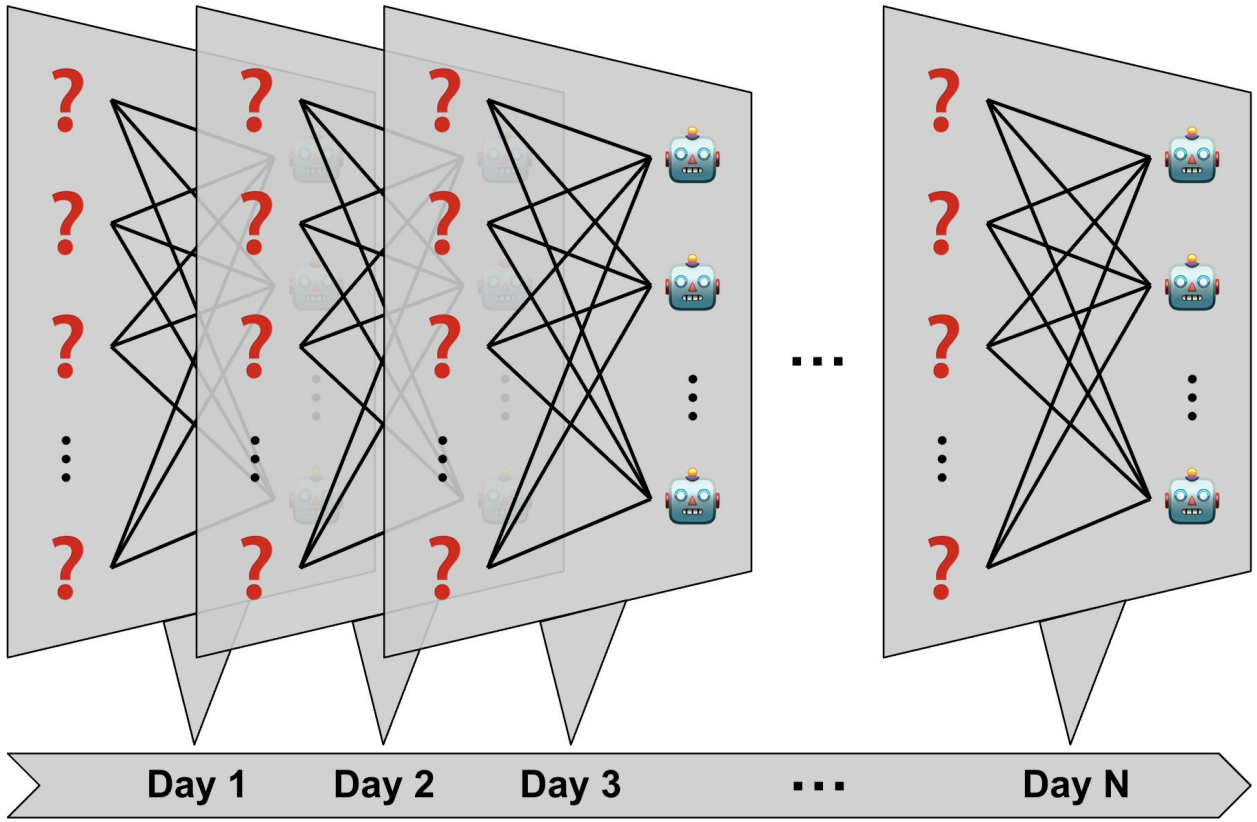


Ask Me Twice

WADL 2026 · October 13, 2026 · Dallas, TX



# Ask Me Twice: a daily archive of chatbot answers




**~400K**
responses preserved

**137**
curated questions

**43**
models tracked

**~5K**
added per day

Coverage from July 18, 2026. Public access: Explorer, REST API, daily bulk exports, OpenAPI, and MCP.

https://wayback-api.archive.org/services/amt/

Figures observed October 12, 2026

# Not every question needs the web, but some really do

| | | |
|---|---|---|
| **13** | **Search required**<br>The answer turns on the present state of the world | *“What are the 10 biggest news stories from today?”* |
| **70** | **Search helps**<br>A factual backbone that moves underneath | *“Is the US prepared to deal with another pandemic?”* |
| **47** | **Answerable from memory**<br>Normative, closed, or about the model itself | *“Should the death penalty be abolished?”* |

Classification of the 130 questions asked on September 11, 2026.

# Asked for today’s news without search: four patterns

### Said so plainly

`claude-sonnet-5 · 117 words`

*“I don’t have access to real-time information, including today’s news. My knowledge has a cutoff date, and I can’t browse the internet… so I genuinely don’t know what’s happening today.”*

### Sent the reader elsewhere

`gpt-5 · 175 words`

*“I don’t have live access to today’s news… Check wire services first: Reuters Top News, AP Top Stories…” — then nine more outlets to go and read instead.*

### Promised, then stopped

`grok-4.5 · 15 words`

*“Gathering today’s top headlines. I’ll pull the latest top stories from major sources right now.” — and then nothing.*

### Invented the news

`o3 · 603 words`

Hedged once at the top, then filed a detailed and entirely fictional bulletin.

*Only the first is honest. Three of the thirty-seven could search that day; the rest answered like this.*

# What a model writes when it cannot check

openai/o3, asked for the day's biggest stories. It noted it was "not connected to live news feeds", then wrote six hundred words of specific, confident reporting.

> "Hurricane Iris made landfall near Galveston, Texas as a Category 4, the strongest September storm since 2005. Early counts: at least 38 dead, $40 billion in insured losses…"

> "The court has agreed to hear Bell v. Schema, the first major case that could narrow tech platforms' liability shield…"

## There is no Hurricane Iris. There is no Bell v. Schema.

The death toll, the dollar figure, and the case name were generated to fit the shape of a news bulletin. A reader skimming past the one-line caveat has no way to tell this from the real thing.

# With search, five of six led with the same story

The same question on September 11, 2026. Four providers, five models, one lead: the twenty-fifth anniversary of the September 11 attacks.

| | |
|---|---|
| `anthropic/claude-sonnet-5` | *“25th Anniversary of 9/11 — Today marks a quarter-century since the September 11, 2001 attacks…”* |
| `google/gemini-3.5-flash` | *“The 25th Anniversary of 9/11 is Commemorated Across the U.S.”* |
| `meta/muse-spark-1.2` | *“United States Marks 25th Anniversary of 9/11”* |
| `openai/gpt-5` | *“9/11 at 25 years: Nationwide commemorations and fresh looks at the attacks’ legacy…”* |
| `openai/o3` | *“25th-anniversary observances of the 9/11 attacks”* |

*Cross-model agreement was not previously available — there was nothing to agree about.*

The sixth, grok-4.5, returned a bare list of links dated to the previous day.

# Enabling search gave the answers citations

*“What are the 10 biggest news stories from today?”*

25th anniversary of the September 11 attacks is marked across the United States, with commemorations at the **[1]** memorial sites and fresh reporting on the attacks’ legacy **[2]** …

[1] apnews.com/article/sept-11-anniversary…
[2] reuters.com/world/us/9-11-twenty-five…
[3] en.wikipedia.org/wiki/September_11_attacks

## Ordinary web resources

- Subject to the same decay as any other
- Outside AMT’s control
- The only part of the answer a reader can verify

# AMT preserves the claim, not the evidence

**Preserved**

Answer text, model, date, raw payload — immutable

**Not preserved**

Every page the answer rests on — live web

*Once cited sources have moved, changed, or been removed, what remains is an assertion that cannot be audited.*

# Study design: one change, two consecutive days

September 11, 2026: provider-native web search enabled for models whose catalog entry established the capability — an index the provider chose, not one substituted at routing.

**137**
questions, unchanged

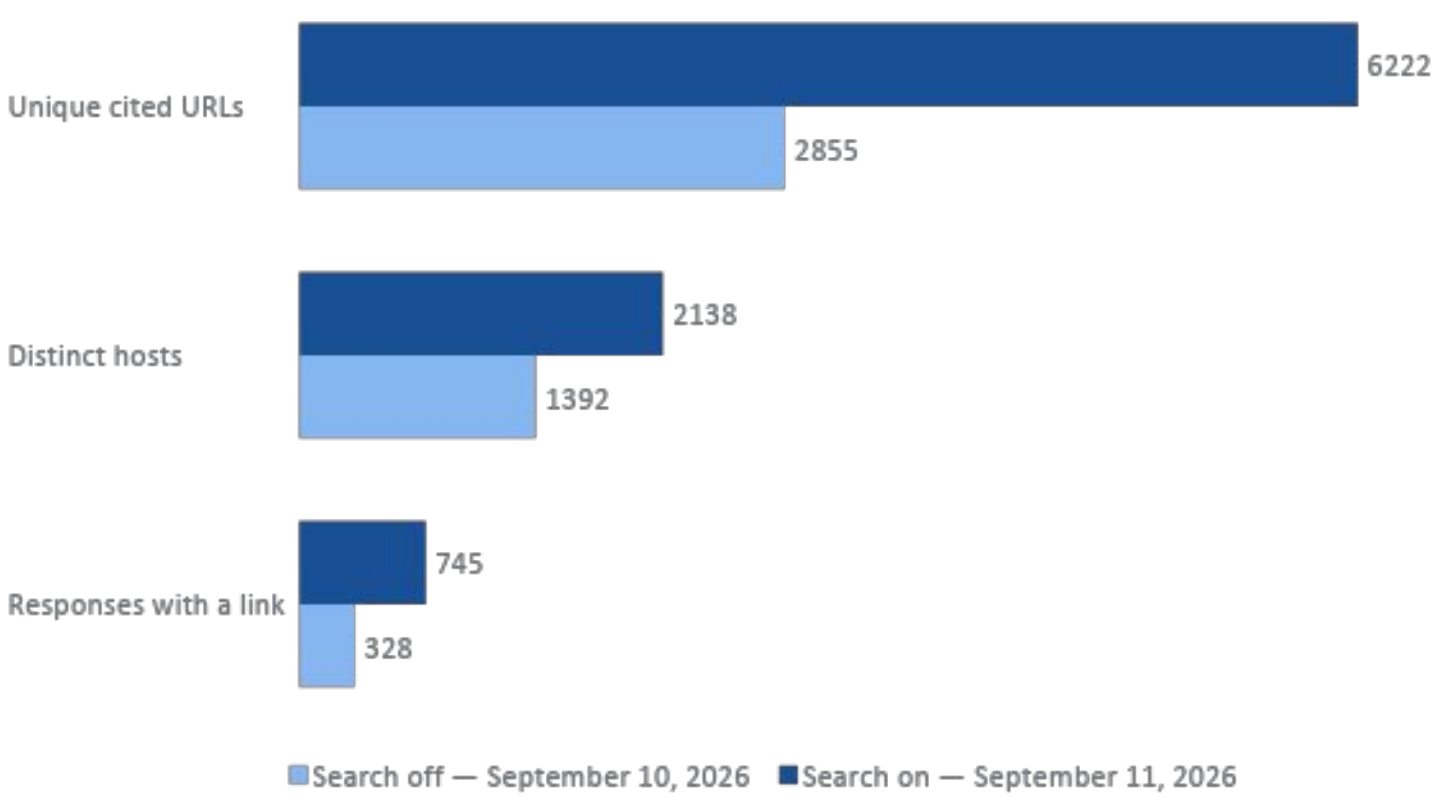
**37**
models, unchanged

**4,810**
responses each day

**6 + 2**
gained search / already had it

- Two Perplexity models searched on every request in both conditions — unchanged control
- Daily cost $20.04 → $54.84, entirely within the six affected models

# Enabling search doubled the cited-web footprint




**2.2×**

more unique URLs to preserve

**2 → 8**

models emitting structured references

*Same 4,810 responses. What AMT depends on grew, not AMT itself.*

# Models deliver URLs two ways, and both must be read

### Structured reference

A url_citation object beside the answer. Machine-readable, directly extractable.

**5,797 URLs · 93.2%**

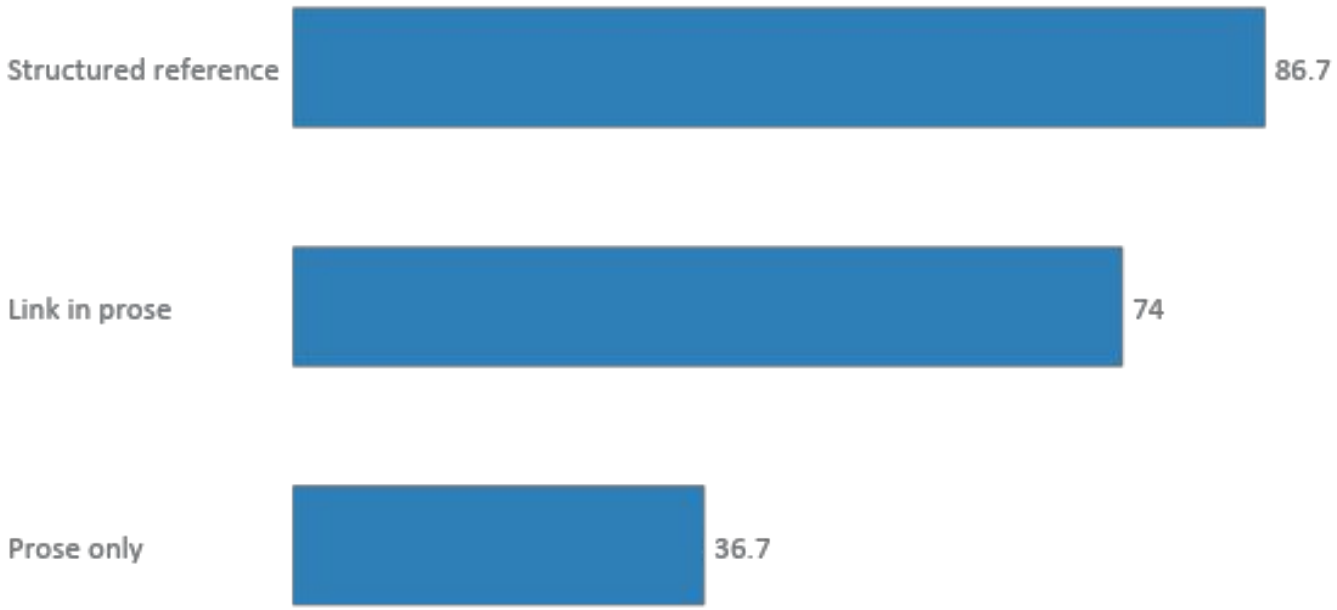
### Written into the prose

A link typed into the answer text. Invisible to an extractor reading only the structured objects.

**1,820 URLs · 29.3%**

- 425 URLs (6.8%) appeared only in prose
- 256 of those came from nine models with no search capability, recalled from training data

# Wayback coverage is lowest for prose-only citations




**50**

percentage point gap between structured references and prose-only citations

*Extraction must read the answer text, not only the metadata.*

# Interstitial hosts obscure the cited publisher

**`vertexaisearch.cloud.google.com`**

613 URLs · 9.8% of the day

**0.0%**

had any capture at the citation date

**1**

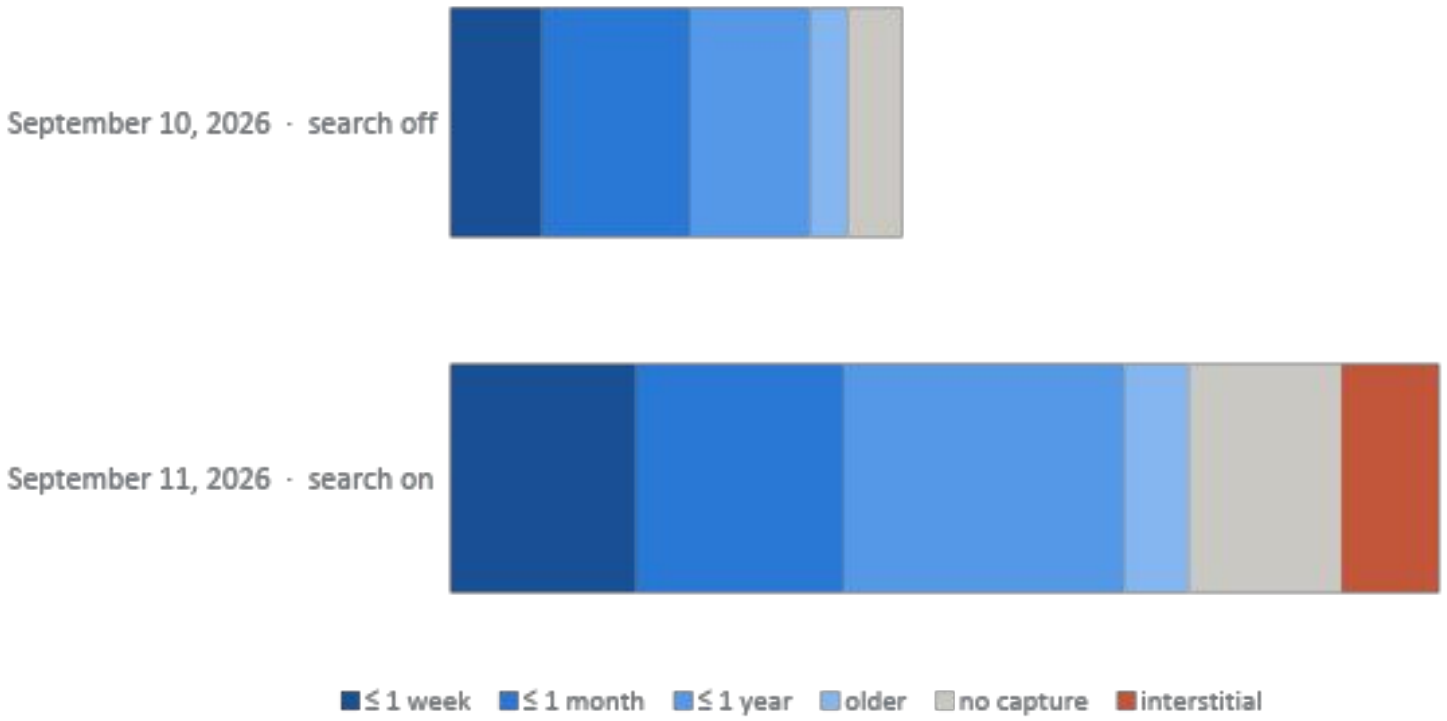
model produced essentially all of them

**1,200+**

named publishers cited by Perplexity at comparable volume

- Not unarchivable — Save Page Now follows 3xx, capturing interstitial and target
- But perishable: redirect services often stop answering before the resources behind them

# Coverage declined as the footprint grew




**87.3%**

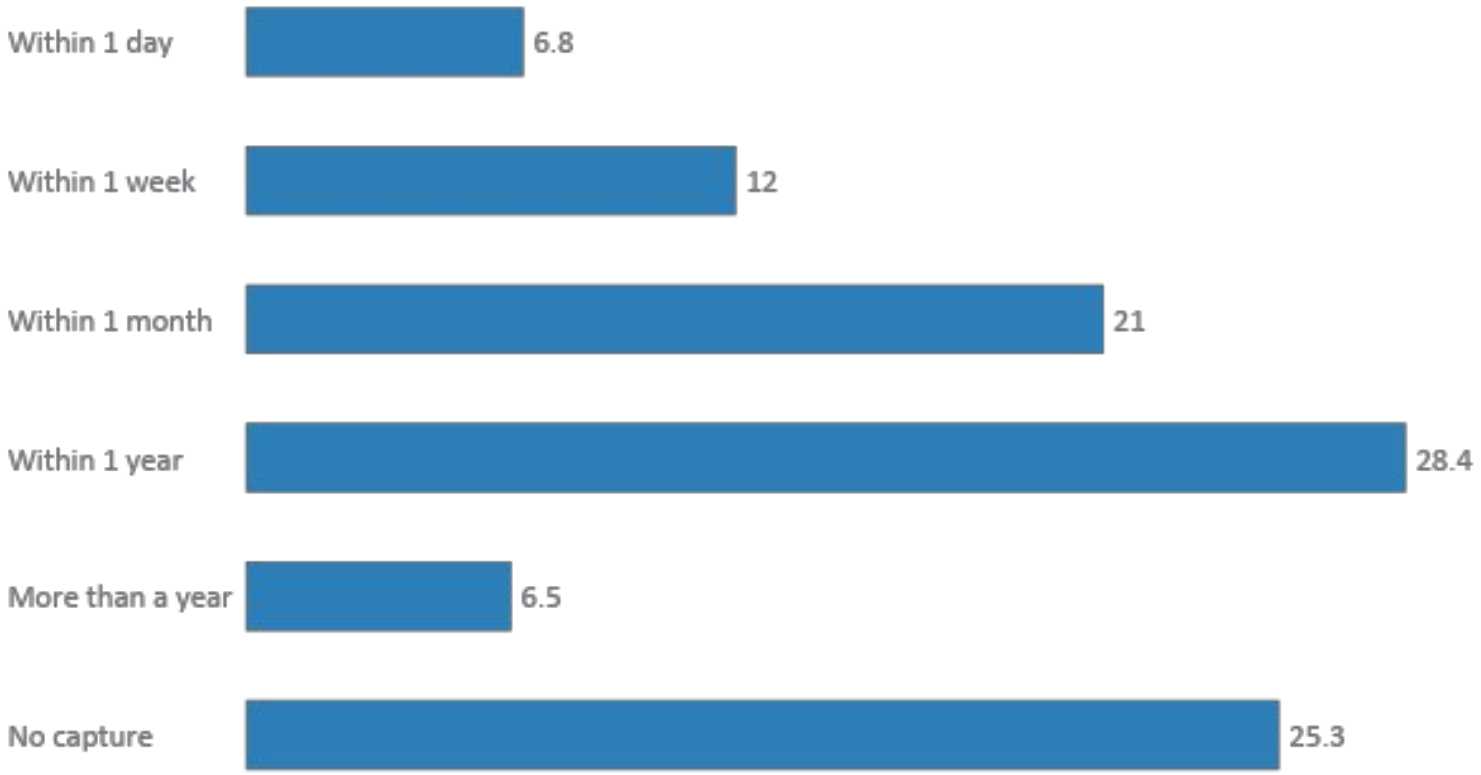
archived, search off

**74.7%**

archived, search on

*Twice as much to preserve, a smaller share already preserved.*

# Most captures long predate the citation




**6.8%**

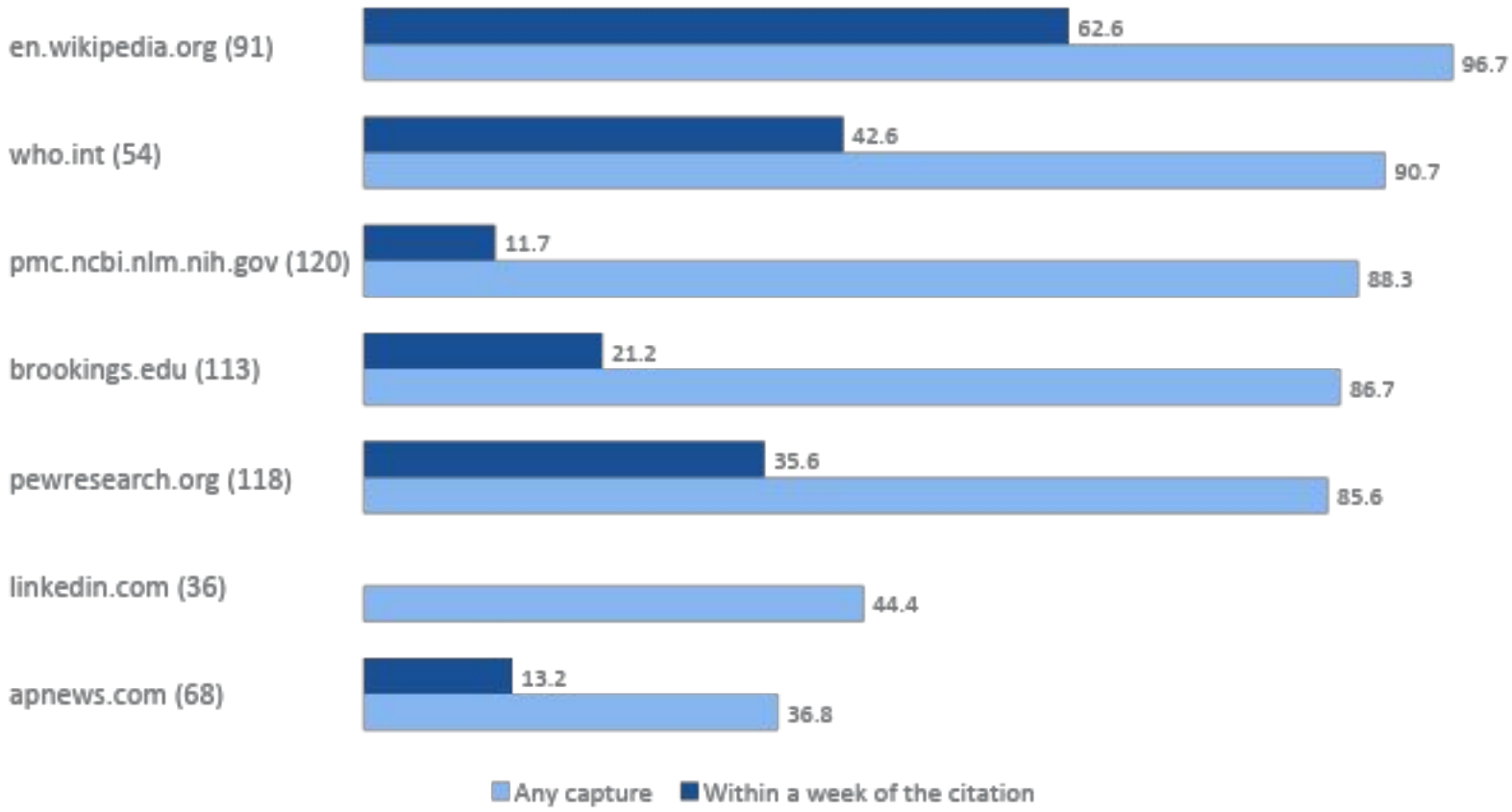
captured within a day of being cited

**87.1%**

of existing captures predate the citation

*Content drift, not absence, is the common failure.*

# Coverage by host: any capture versus one within a week




Share of each host's cited URLs. Counts in brackets are the URLs cited on that host that day.

**11.7%**

of cited PubMed Central URLs have a capture within a week — against 88.3% with any capture

**1,482**

of the day's 2,131 distinct hosts were cited exactly once

*A curated seed list would omit most cited hosts.*

# Pipeline




- Extract every link from the answer text — structured references and prose alike
- Canonicalize, reject what only looks URL-shaped, and dedupe per response and per day
- Submit in rate-limited batches, resumable from a checkpoint

*Two scripts and an existing preservation service.*

# Throughput and cost of a 1,527-URL run

**65.6%**
captured successfully

**3h 34m**
to resolve 1,527 URLs

**7.1**
resolutions per minute

**~14h**
projected for a full day

- 7.1/min is the authenticated submission ceiling, not a limit of the tool
- Submission is asynchronous; elapsed time is polling to resolution, which varies with load
- Median capture 15.4 s, p90 21.1 s

*Tractable as an overnight batch. Precludes synchronous capture during collection.*

# Failure modes: refusal vs. decay

**389**

**HTTP 403**

server refused the archiver — 74% of 526 failures

**73**

**HTTP 404**

not found — the canonical link-rot case, 14%

**<20**

**Transient**

rate limiting, gateway errors

**194**

**Hosts**

returned 403 — broad, not isolated

*Access policy, more than resource decay, limits what can be preserved — and it falls hardest on the authoritative sources citations are meant to invoke.*

Wikipedia accounted for 35 of these 403s and was refusing Save Page Now only at the time of measurement; it no longer does. Excluding it, refusals are still 354 of 526.

# Implications for archives of generated content

1

## Treat citation outlinks as in scope

A response archive that stops at the answer text preserves a claim that becomes progressively less verifiable.

2

## Extract from the text, not only the metadata

Structured references are convenient and incomplete. The shortcut fails where it matters most.

3

## Citation form determines urgency

A citation routed through a redirect conceals its source until dereferenced, and the redirect may stop answering before the resource does.

# Future work

- Citation submission folded into the nightly schedule
- Beyond the single day pair — the full collection period
- Dereference interstitials to recover true publisher distributions
- Separate access-policy refusals from archiver-identity effects
- Responses and their captured citations exported together as WARC
- Second routing provider, to test how much is OpenRouter-specific

*Direction is robust and the control unchanged; magnitudes remain provisional on two days.*

# Conclusions

*Enabling search gave AMT an evidence base, and an obligation toward it.*

### Citations are archive scope

An answer whose sources are gone cannot be audited

### Read the text, not the metadata

6.8% of URLs are prose-only — archived at 36.7% against 86.7%

### Capture at citation time

87.1% of existing captures predate the citation they support

### Refusal outweighs decay

403s outnumber 404s roughly five to one, across 194 hosts

### Citation form decides urgency

Redirects resolve today, and may stop before the page does

### The method is ordinary

Two scripts and Save Page Now; the obligation is the finding

**Ask Me Twice** — Explorer, API, daily exports, and MCP: **https://wayback-api.archive.org/services/amt/**