SECTION 6 · 15 min

Temporal Web & Evaluation Ecosystem

Connecting temporal IR to the evolving Web and the community resources used to evaluate time-aware systems.

Presenter: Bhawna Piryani

06

Dynamic Web vs. Archived Web

We search the present on the live Web, and the past in Web archives.

Dynamic Web

Current, live, constantly changing

Characteristics

- Content created and updated continuously
- Freshness is a key relevance signal
- Search indexes the latest versions

Typical queries

"Latest Apple earnings"
"GPT-4o release date"
"Breaking news on inflation"

Archived Web

Historical, preserved snapshots

Characteristics

- Preserves past versions of pages
- Access to information as it existed
- Provided by Web archives (Internet Archive)

Typical queries

"CDC website on COVID, Mar 2020"
"How was the iPhone 5s announced?"
"Iraq war coverage in 2003"

Link Rot & Content Drift

Two different problems — and the difference matters whenever you cite the Web.

Link rot

The page is gone — the URL returns 404 and the reference can't be reached.

`example.org/report → 404 Not Found`

Content drift

The page still loads, but its content changed — so it no longer matches what was cited.

`apple.com 2007 → 2026 (different page)`

Problem	URL works?	Content same?
Link rot	✗	✗
Content drift	✓	✗

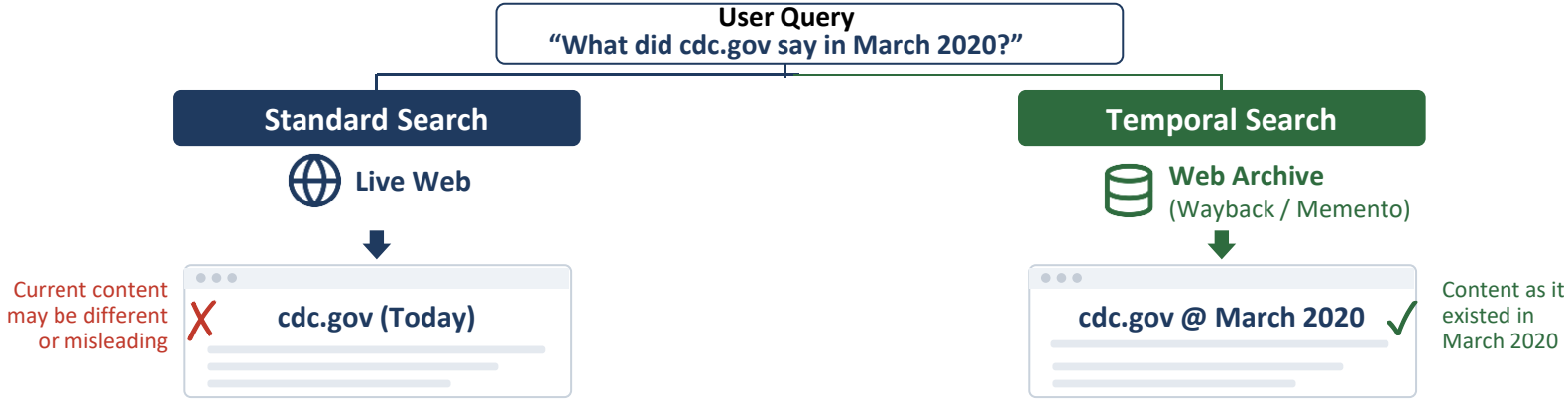
Drift is the dangerous case: the link still works, so the change slips by unnoticed.

Time-Travel Search

Retrieving the Web as it existed at a specific point in time.

User Question:
What did the CDC website say about COVID-19 in March 2020?

Standard Web Search	Temporal Web Search
Searches the live Web	Searches the archived Web
Retrieves the current version of cdc.gov	Retrieves the March 2020 snapshot of cdc.gov
Answers reflect the most recent information	Answers reflect the information at that time
May be incorrect for historical questions ✗	Provides the correct historical state ✓



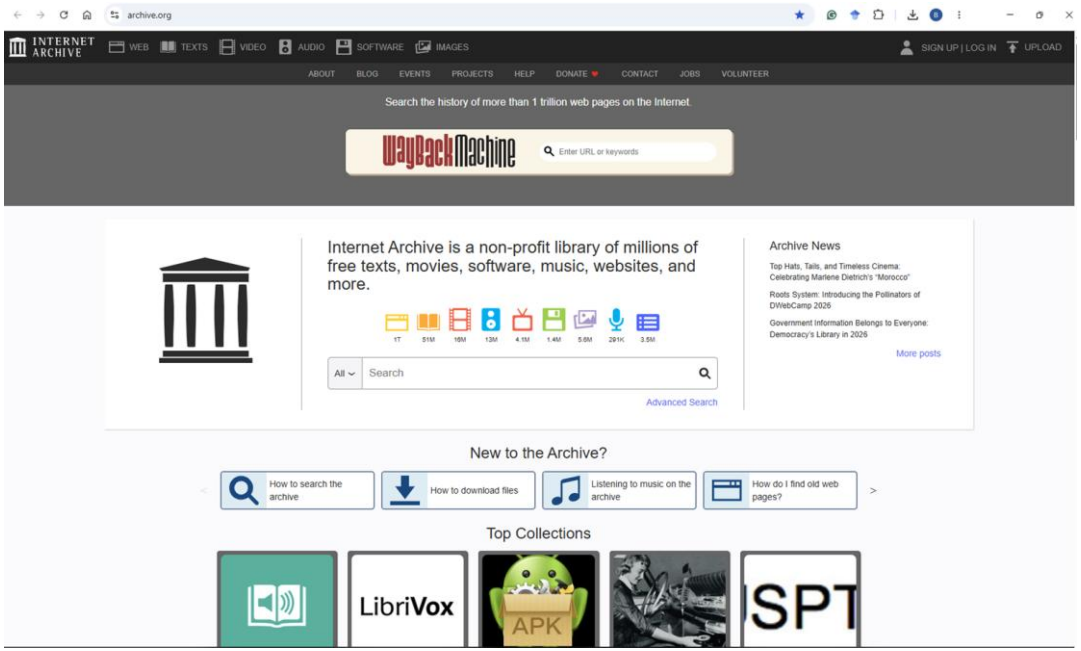
The Internet Archive

A digital library that preserves historical versions of online and offline digital resources.

What it preserves?

- Web pages and websites
- Books and documents
- Audio, video and software
- Research and cultural collections

The Internet Archive is the preservation infrastructure; the Wayback Machine is its Web-access service.

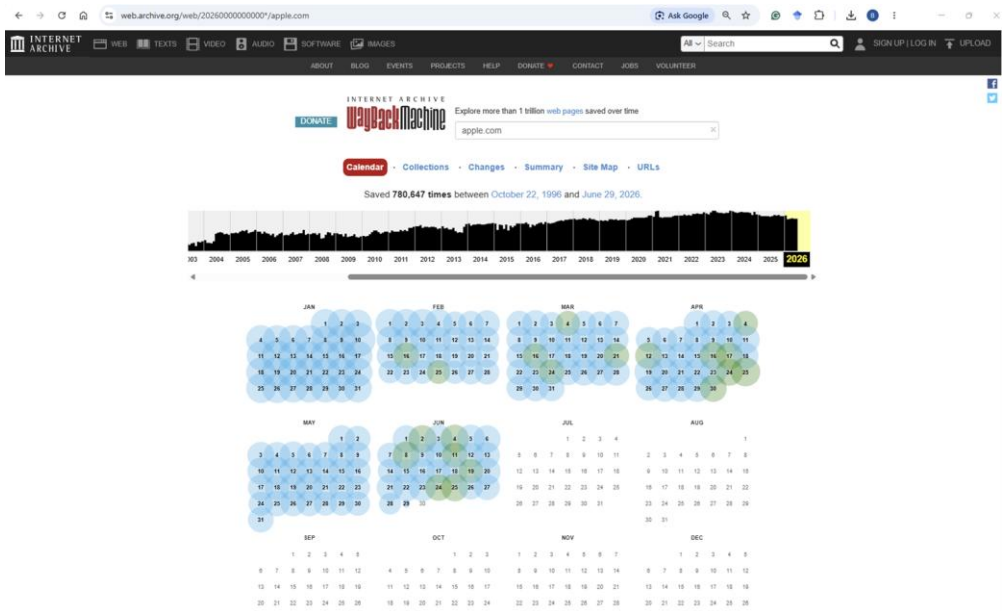


Wayback Machine

Access earlier versions of webpages

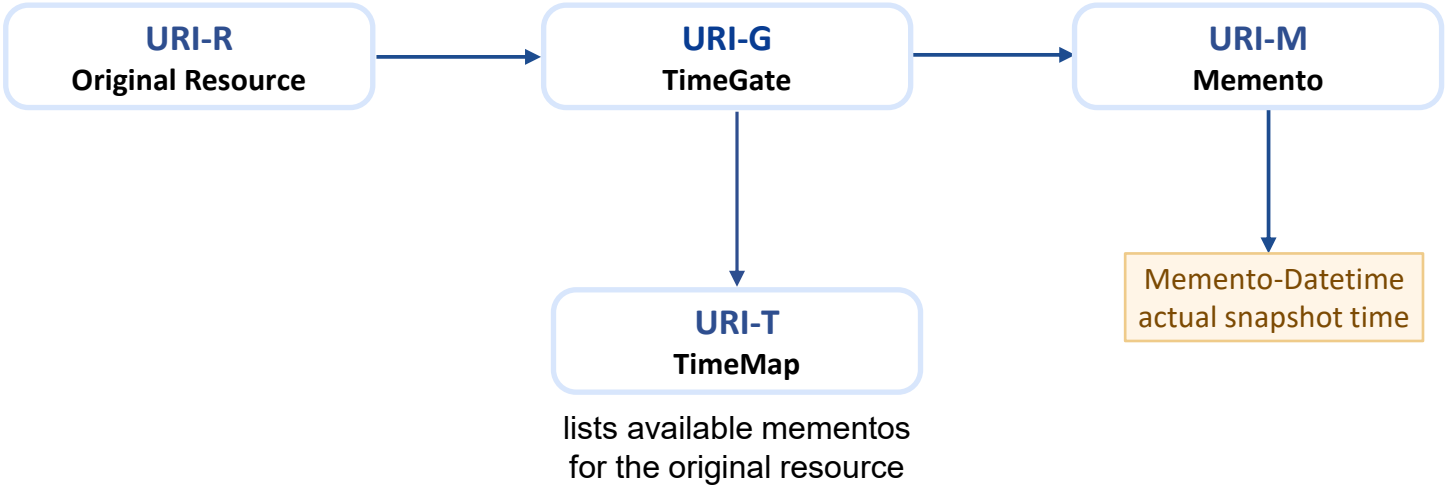
1. Enter the webpage URL
2. Select a year and capture date
3. Open the archived snapshot

URL + timestamp → historical webpage



Memento: HTTP-Based Time Travel for the Web

Memento adds datetime negotiation to HTTP: instead of asking only “which URL?”, the client can ask “which version of this URL near this datetime?”



Memento gives a protocol-level way to discover and access prior versions across web archives, not just through one archive interface.

H. Van de Sompel, M. Nelson, and R. Sanderson. 2013. RFC 7089: **HTTP Framework for Time-Based Access to Resource States -- Memento**. RFC Editor, USA.

WARC: The Storage Format Behind Web Archives

WARC files package captured web resources into an archival container: each record has headers plus the recorded payload.

WARC file

Response
HTML / image / CSS / JS payload

Request
HTTP request context

Metadata
crawl and provenance metadata

Revisit
duplicate or unchanged content reference

Storage vs. access

WARC separates archival storage from replay. Indexes and replay software make captures browsable later.

Why it matters

A WARC record preserves both the captured payload and useful context such as request, metadata, and provenance.

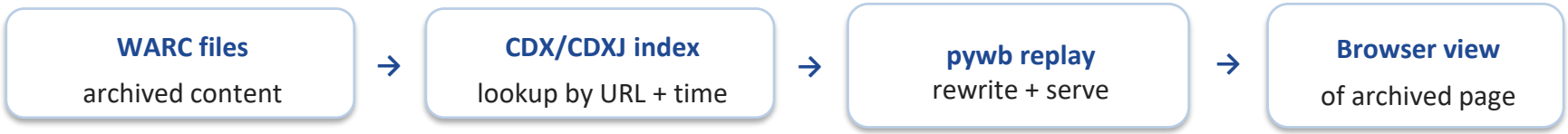
Temporal IR view

The same URL can have many WARC captures at different timestamps.

https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1-annotated/

pywb: Replay and Access for Web Archives

pywb is a Python web archiving toolkit for capture and replay. It can provide basic “Wayback Machine” functionality for your own WARC collections.



- Reproduce experiments on fixed snapshots
- Replay archived pages with rewritten links
- Build controlled temporal collections

Wayback is the large public service; pywb is the reusable software layer that lets researchers build and replay smaller web archives locally.

https://pypi.python.org/pypi/pywb

The Impact of AI-Generated Text on the Internet

Research question

How much newly published Web content is AI-generated or AI-assisted, and how is it changing online discourse?

Data

- Representative samples of websites from the Internet Archive
- Monthly snapshots from August 2022 to May 2025
- Oldest archived version retrieved through the Wayback Machine

Method

Archived webpages → visible text extraction → AI-text detection → longitudinal analysis

Why this case study matters

Web archives make it possible to measure how the composition of the Web changes over time.

Dolezal, Jonas, Sawood Alam, Mark Graham, and Maty Bohacek. "**The Impact of AI-Generated Text on the Internet**." arXiv preprint arXiv:2604.26965 (2026).

The Impact of AI-Generated Text on the Internet

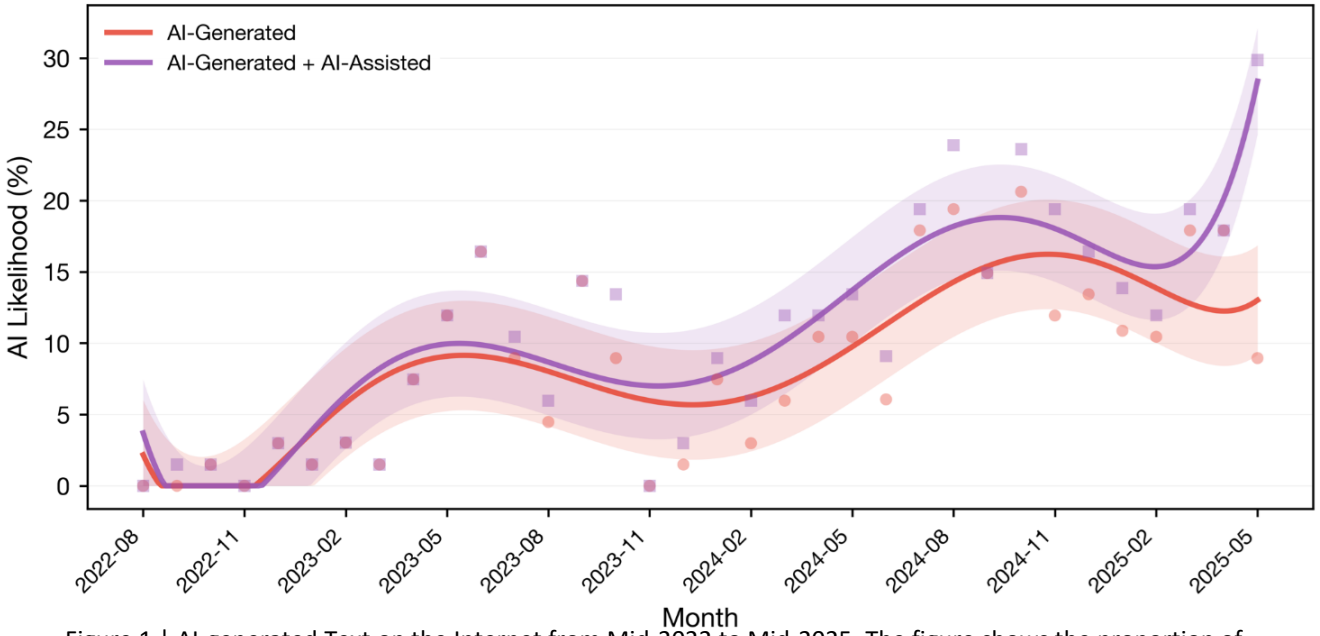


Figure 1 | AI-generated Text on the Internet from Mid-2022 to Mid-2025. The figure shows the proportion of websites classified as fully AI-generated (red) and AI-generated or AI-assisted (purple)based on Pangram v3 detection applied to representative samples obtained from the Internet Archive. Curves represent smoothed estimates.

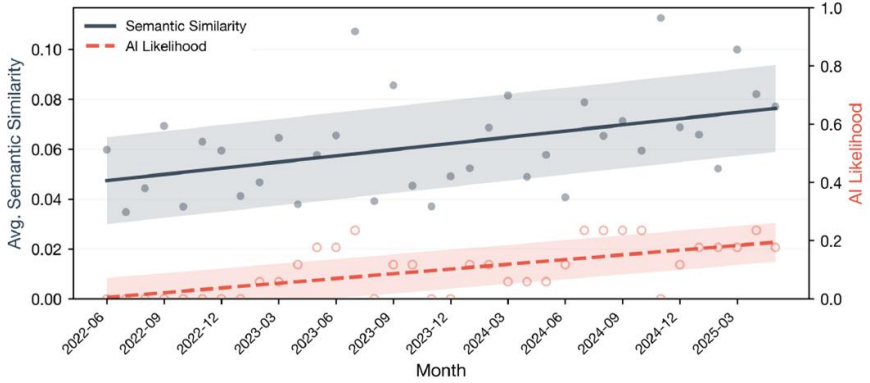
~35%

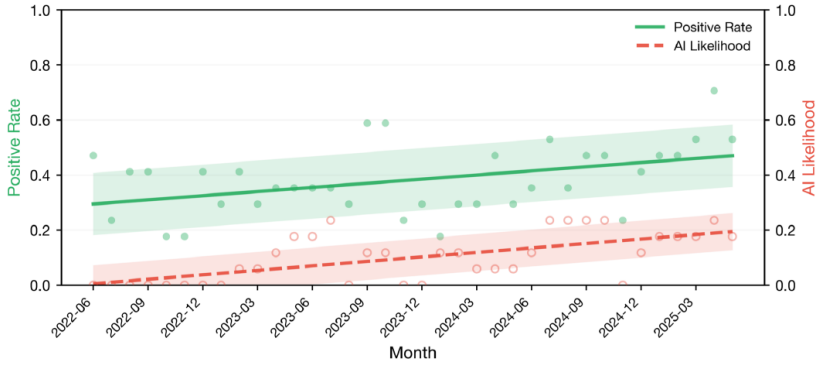
newly published websites classified as AI-generated or AI-assisted by mid-2025.

growth begins after ChatGPT's public launch in late 2022. Before ChatGPT's launch in late 2022, the estimate was close to zero. The rate then increased over time, particularly during 2024 and 2025.

Dolezal, Jonas, Sawood Alam, Mark Graham, and Maty Bohacek. "**The Impact of AI-Generated Text on the Internet**." arXiv preprint arXiv:2604.26965 (2026).

The Impact of AI-Generated Text on the Internet





Semantic Contraction

- As AI likelihood rises, average semantic similarity also rises.
- AI-generated/assisted websites show 33% higher semantic similarity than non-AI websites.
- Interpretation: AI text may narrow the range of meanings or viewpoints online.

Positivity shift

- As AI likelihood rises, the share of positive-sentiment documents also rises.
- Average positive sentiment is 107% higher for AI-generated/assisted websites.
- Interpretation: web discourse may become more cheerful, sanitized, or sycophantic.

Dolezal, Jonas, Sawood Alam, Mark Graham, and Maty Bohacek. "**The Impact of AI-Generated Text on the Internet**." arXiv preprint arXiv:2604.26965 (2026).

The Impact of AI-Generated Text on the Internet

Main takeaways

- AI-generated/assisted text is estimated to be a large and growing share of newly published websites.
- The strongest measured effects are semantic contraction and positivity shift.
- The paper finds a gap between public belief and web-scale evidence for factual/style degradation.

“AI text may not simply make the web less factual; it may make the web more similar, more positive, and harder to interpret as human discourse.”

Dolezal, Jonas, Sawood Alam, Mark Graham, and Maty Bohacek. "**The Impact of AI-Generated Text on the Internet**." arXiv preprint arXiv:2604.26965 (2026).