

# An Initial Test of AI for Temporal Web Forensics

Nicholas Taylor  
[nullhandle.org](https://nullhandle.org)

[AI and Web Archiving Workshop](#): [AI4LAM](#) Affiliate Event  
15 September 2026

"Brighton Clock Tower - festive" by David Pearson under [CC BY-NC-SA 2.0](#)

# Temporal Web Forensics

In my capacity as an expert witness, I'm called upon to investigate and testify regarding **the temporal dimensions of web content** – when it may have been published, updated, available, or discoverable.

My shorthand for this expertise is *temporal web forensics*.

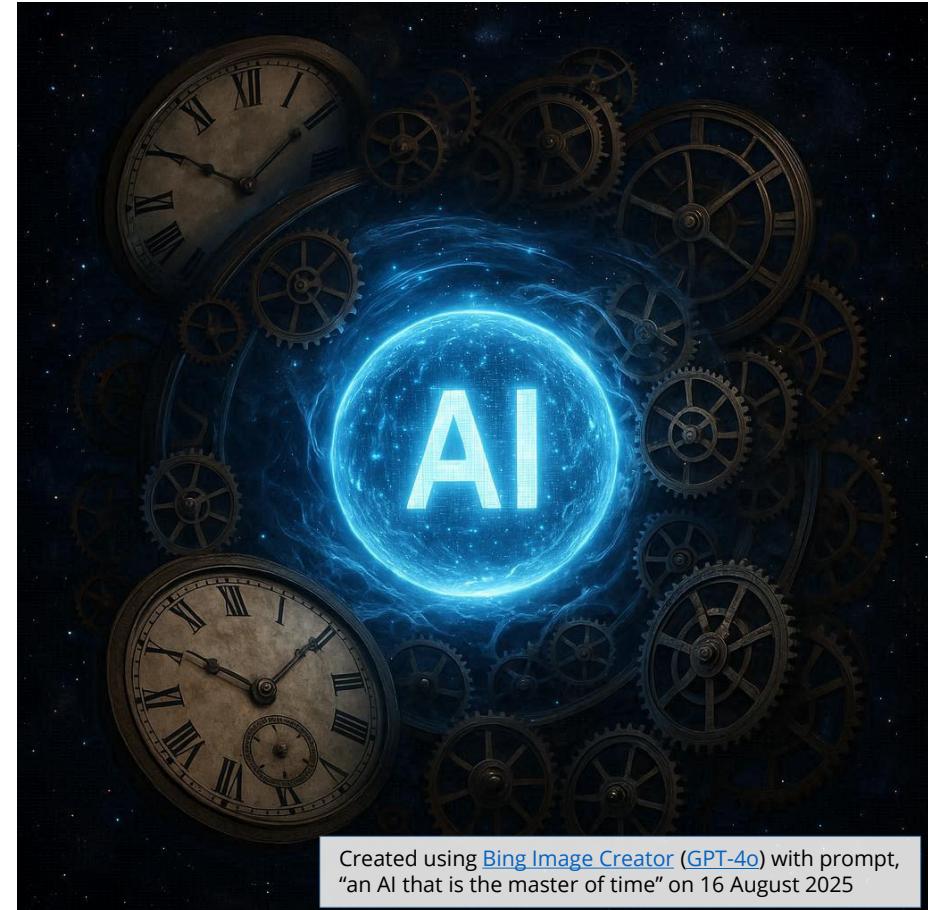


"clock world" by Mathanki Kodavasal under CC BY-SA 2.0

# AI for Temporal Web Forensics?

I wanted to see how capably and reliably LLMs could discern critical temporal dimensions of web content.

So, about a year ago (August 2025), I ran and [documented](#) some tests.



Created using [Bing Image Creator \(GPT-4o\)](#) with prompt, "an AI that is the master of time" on 16 August 2025

# Test Webpages / Websites

- [A SCOTUSBlog post](#)
- [The U.S. Supreme Court website](#)
- [A webpage on the U.S. Supreme Court website featuring a speech by Chief Justice Rehnquist](#)
- [The website for the 1996 film \*Space Jam\*](#)
- [A post from my own blog](#)

# Experimental Setup

## Tests

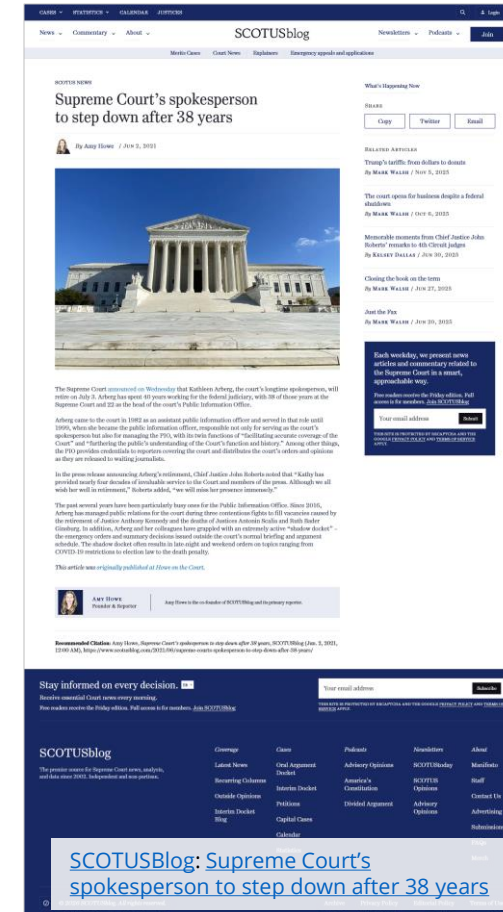
- [Zero-shot](#): asking when the website or webpage was published
- [Iterative](#): if zero-shot yielded a poor answer, follow up
- Screenshot interpretation: estimate publication date based on screenshot

## Models

- [ChatGPT 4o](#)
- [Gemini 2.5 Flash](#)
- [Grok 3](#)
- [Sonnet 4](#)

# SCOTUSBlog Post

- Most straight-forward case
- Likely published **2 June 2021**, corroborated by:
  - [Body text timestamp](#)
  - [dateCreated](#) attribute
  - [Earliest IAWM capture](#)
  - Google [bylineDate](#)



# SCOTUSBlog Post Test

| Model   | Version   | Query type                | Date in response       |
|---------|-----------|---------------------------|------------------------|
| ChatGPT | GPT-4o    | Zero-shot                 | 2021-06-02             |
| ChatGPT | GPT-4o    | Screenshot interpretation | 2021-06-02             |
| Gemini  | 2.5 Flash | Zero-shot                 | 2021-06-02             |
| Gemini  | 2.5 Flash | Screenshot interpretation | Late 2020 / early 2021 |
| Grok    | 3         | Zero-shot                 | 2021-06-29             |
| Grok    | 3         | Iterative                 | 2021-06-29             |
| Grok    | 3         | Screenshot interpretation | 2021-06-02             |
| Sonnet  | 4         | Zero-shot                 | 2021-06-02             |
| Sonnet  | 4         | Screenshot interpretation | 2021-06-02             |

# U.S. Supreme Court Website

- Can determine approximate publication date
- [WHOIS domain registration](#):  
**1 December 1997**
- [Earliest IAWM capture](#):  
**20 May 2000**
  - [x-archive-orig-last-modified](#)  
HTTP header: **27 April 2000**



# U.S. Supreme Court Website Test

| Model   | Version   | Query type                | Date in response         |
|---------|-----------|---------------------------|--------------------------|
| ChatGPT | GPT-4o    | Zero-shot                 | 2000-04-17               |
| ChatGPT | GPT-4o    | Screenshot interpretation | Late 1990s / early 2000s |
| Gemini  | 2.5 Flash | Zero-shot                 | Can't say                |
| Gemini  | 2.5 Flash | Iterative                 | 2000-04-17               |
| Gemini  | 2.5 Flash | Screenshot interpretation | Late 1990s / early 2000s |
| Grok    | 3         | Zero-shot                 | 2000-04                  |
| Grok    | 3         | Screenshot interpretation | 1997 – 2002              |
| Sonnet  | 4         | Zero-shot                 | Mid-to-late 1990s        |
| Sonnet  | 4         | Iterative                 | 1996 – 1998              |
| Sonnet  | 4         | Screenshot interpretation | Late 1990s / early 2000s |

# Speech by Chief Justice Rehnquist

- Many contradictory dates
- Body text: **3 May 2000**
- <meta> created + revised:  
**31 December 1600**
- Google bylineDate: **5 March 2000**
- Earliest IAWM capture (new domain): **30 April 2010**
- Earliest IAWM capture (old domain): **16 August 2000**
  - x-archive-orig-last-modified HTTP header: **16 May 2000**



SUPREME COURT  
OF THE UNITED STATES

[ADVANCED SEARCH](#)  
[DOCKET SEARCH](#)

SEARCH TIPS

[OPINIONS](#)
[FILING & RULES](#)
[ORAL ARGUMENTS](#)
[CASE DOCUMENTS](#)
[NEWS MEDIA](#)
[ABOUT](#)
[VISIT](#)



[HOME](#) > [NEWS MEDIA](#) > [SPEECHES](#) > [05-03-00](#)

Services for News Media

Press Releases

Media Advisories

Press Credentials

**Speeches**

A Reporter's Guide to Applications

Chief Justice's Year-End Reports on the Federal Judiciary

**EMBARGOED FOR RELEASE**

May 3, 2000

Remarks of Chief Justice William H. Rehnquist  
100th Anniversary Celebration  
Of the Norfolk and Portsmouth Bar Association  
Norfolk, Virginia  
May 3, 2000

Thank you, Judge Doumar, for your kind introduction. I am very pleased to be here today to help celebrate the 100th anniversary of the founding of the Norfolk and Portsmouth Bar Association.

I am going to speak this afternoon about civil liberty in time of war, focusing first on the Civil War and then on World War II. I have chosen this subject in part because of the importance of the Civil War in this historic area.

Even those of you who did not major in history probably know that Abraham Lincoln was elected President in November of 1860, and was inaugurated as President on March 4, 1861. Between the time of his election and his inauguration, the seven states of the deep south -- South Carolina, Georgia, Florida, Alabama, Mississippi, Louisiana, and Texas -- had seceded from the Union and elected Jefferson Davis as their President. For the first six weeks of Lincoln's administration, the cabinet debated what to do about the Union garrison at Fort Sumter, on an island in the harbor of Charleston, South Carolina. In mid-April, the Confederate shore batteries opened up on the fort, and the garrison surrendered the next day. Lincoln called for 75,000 volunteers to put down the rebellion, and the four states of the upper south -- Virginia, North Carolina, Tennessee, and Arkansas -- seceded and joined the original seven states of the Confederacy. The Civil War had begun.

As most of you already know, some of the more well-known Civil War events occurred in this area. Indeed, the Norfolk Navy Shipyard played an interesting role during the Civil War. The U. S. government had established the Norfolk Navy Shipyard in 1801. Its predecessor on the same site was a private shipyard built in 1767 by a wealthy Scottish merchant named Andrew Sprowle.

Supreme Court of the United States: Speeches - sp\_05-03-00

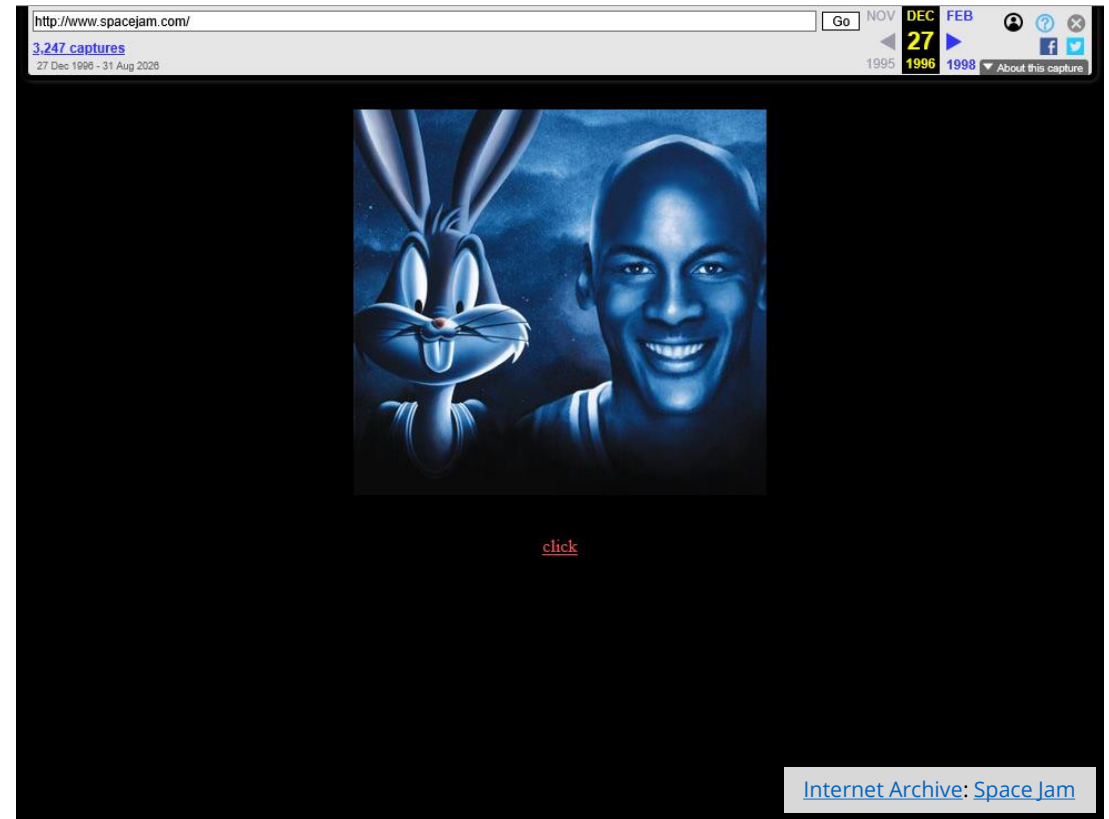
On April 30, 1861, 1860, 1861, 1862, 1863, 1864, 1865, 1866, 1867, 1868, 1869, 1870, 1871, 1872, 1873, 1874, 1875, 1876, 1877, 1878, 1879, 1880, 1881, 1882, 1883, 1884, 1885, 1886, 1887, 1888, 1889, 1890, 1891, 1892, 1893, 1894, 1895, 1896, 1897, 1898, 1899, 1900, 1901, 1902, 1903, 1904, 1905, 1906, 1907, 1908, 1909, 1910, 1911, 1912, 1913, 1914, 1915, 1916, 1917, 1918, 1919, 1920, 1921, 1922, 1923, 1924, 1925, 1926, 1927, 1928, 1929, 1930, 1931, 1932, 1933, 1934, 1935, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973, 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025

# Speech by Chief Justice Rehnquist Test

| Model   | Version                                       | Query type                | Date in response |
|---------|-----------------------------------------------|---------------------------|------------------|
| ChatGPT | GPT-4o                                        | Zero-shot                 | 2000-05-03       |
| ChatGPT | GPT-4o                                        | Screenshot interpretation | 2000-05          |
| Gemini  | 2.5 Flash                                     | Zero-shot                 | 2002-03-21       |
| Gemini  | 2.5 Flash                                     | Iterative                 | 2000-03-21       |
| Gemini  | 2.5 Flash (w/ <a href="#">Deep Research</a> ) | Zero-shot                 | 2000-05-03       |
| Gemini  | 2.5 Flash                                     | Screenshot interpretation | 2000-05-03       |
| Grok    | 3                                             | Zero-shot                 | 2000-05-03       |
| Grok    | 3                                             | Screenshot interpretation | 2020-06-08       |
| Sonnet  | 4                                             | Zero-shot                 | 2000-05-03       |
| Sonnet  | 4                                             | Screenshot interpretation | 1995 – 1999      |

# *Space Jam* Website

- Can determine approximate publication date
- [WHOIS domain registration](#):  
**14 March 1996**
- [Earliest IAWM capture](#):  
**27 December 1996**

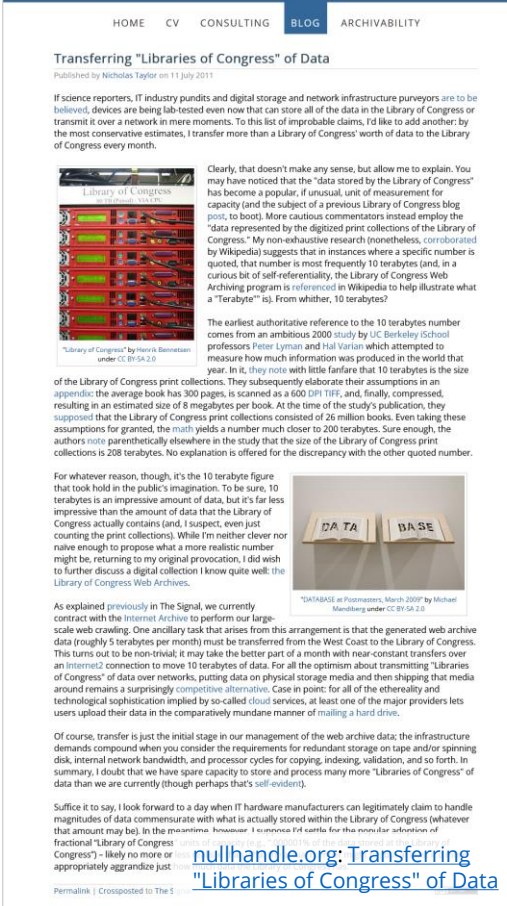


# *Space Jam* Website Test

| Model   | Version   | Query type                | Date in response |
|---------|-----------|---------------------------|------------------|
| ChatGPT | GPT-4o    | Zero-shot                 | 1996             |
| ChatGPT | GPT-4o    | Screenshot interpretation | 1996             |
| Gemini  | 2.5 Flash | Zero-shot                 | 1996             |
| Gemini  | 2.5 Flash | Iterative                 | 1996-11          |
| Gemini  | 2.5 Flash | Screenshot interpretation | 1996             |
| Grok    | 3         | Zero-shot                 | 1996             |
| Grok    | 3         | Screenshot interpretation | 1996             |
| Sonnet  | 4         | Zero-shot                 | 1996             |
| Sonnet  | 4         | Screenshot interpretation | 1996 – 1998      |

# nullhandle.org Blog Post

- Many different dates
- [Body text](#): 11 July 2011
- Web address: 11 July 2011
- [WHOIS domain registration](#): 17 September 2017
- [datePublished](#) attribute: 11 July 2011
- [dateModified](#) attribute: 29 September 2017



The screenshot shows a blog post on the nullhandle.org website. The navigation bar at the top includes links for HOME, CV, CONSULTING, BLOG (which is highlighted), and ARCHIVABILITY. The post title is "Transferring \"Libraries of Congress\" of Data" and it is attributed to Nicholas Taylor, published on 11 July 2011. The main text discusses the challenges of transferring data from the Library of Congress, mentioning the 10 terabyte figure and the need for a more realistic number. It includes two images: one showing a stack of books labeled "Library of Congress" and another showing a book titled "DATABASE as Postmasters, March 2009" by Michael Mandelberg. The post concludes with a summary of the data transfer process and a link to the nullhandle.org website.

HOME CV CONSULTING **BLOG** ARCHIVABILITY

## Transferring "Libraries of Congress" of Data

Published by Nicholas Taylor on 11 July 2011

If science reporters, IT industry pundits and digital storage and network infrastructure purveyors are to be believed, devices are being lab-tested even now that can store all of the data in the Library of Congress or transmit it over a network in mere moments. To this list of improbable claims, I'd like to add another: by the most conservative estimates, I transfer more than a Library of Congress' worth of data to the Library of Congress every month.

Clearly, that doesn't make any sense, but allow me to explain. You may have noticed that the "data stored by the Library of Congress" has become a popular, if unusual, unit of measurement for capacity (and the subject of a previous Library of Congress blog post, to boot). More cautious commentators instead employ the "data represented by the digitized print collections of the Library of Congress." My non-exhaustive research (nonetheless, corroborated by Wikipedia) suggests that in instances where a specific number is quoted, that number is most frequently 10 terabytes (and, in a curious bit of self-referentiality, the Library of Congress Web Archiving program is referenced in Wikipedia to help illustrate what a "terabyte" is). From whither, 10 terabytes?

The earliest authoritative reference to the 10 terabytes number comes from an ambitious 2000 study by UK Berkeley School professors Peter Lyman and Hal Varian which attempted to measure how much information was produced in the world that year. In it, they note with little fanfare that 10 terabytes is the size of the Library of Congress print collections. They subsequently elaborate their assumptions in an appendix: the average book has 300 pages, is scanned as a 600 DPI TIFF, and, finally, compressed, resulting in an estimated size of 8 megabytes per book. At the time of the study's publication, they supposed that the Library of Congress print collections consisted of 26 million books. Even taking these assumptions for granted, the math yields a number much closer to 200 terabytes. Sure enough, the authors note parenthetically elsewhere in the study that the size of the Library of Congress print collections is 208 terabytes. No explanation is offered for the discrepancy with the other quoted number.

For whatever reason, though, it's the 10 terabyte figure that took hold in the public's imagination. To be sure, 10 terabytes is an impressive amount of data, but it's far less impressive than the amount of data that the Library of Congress actually contains (and, I suspect, even just counting the print collections). While I'm neither clever nor naive enough to propose what a more realistic number might be, returning to my original provocation, I did wish to further discuss a digital collection I know quite well: the Library of Congress Web Archives.

As explained previously in The Signal, we currently contract with the Internet Archive to perform our large-scale web crawling. One ancillary task that arises from this arrangement is that the generated web archive data (roughly 5 terabytes per month) must be transferred from the West Coast to the Library of Congress. This turns out to be non-trivial: it may take the better part of a month with near-constant transfers over an Internet2 connection to move 10 terabytes of data. For all the optimism about transmitting "Libraries of Congress" of data over networks, putting data on physical storage media and then shipping that media around remains a surprisingly competitive alternative. Case in point: for all of the etherality and technological sophistication implied by so-called cloud services, at least one of the major providers lets users upload their data in the comparatively mundane manner of mailing a hard drive.

Of course, transfer is just the initial stage in our management of the web archive data: the infrastructure demands compound when you consider the requirements for redundant storage on tape and/or spinning disk, internal network bandwidth, and processor cycles for copying, indexing, validation, and so forth. In summary, I doubt that we have spare capacity to store and process many more "Libraries of Congress" of data than we are currently (though perhaps that's self-evident).

Suffice it to say, I look forward to a day when IT hardware manufacturers can legitimately claim to handle magnitudes of data commensurate with what is actually stored within the Library of Congress (whatever that amount may be). In the meantime, however, I suppose I'll settle for the non-trivial selection of fractional "Library of Congress" data to transfer to the Library of Congress. I'll leave the rest of the data to the Internet Archive.

For more information, see the nullhandle.org website: [nullhandle.org: Transferring "Libraries of Congress" of Data](#)

Permalink | Crossposted to The Signal

# nullhandle.org Blog Post Test

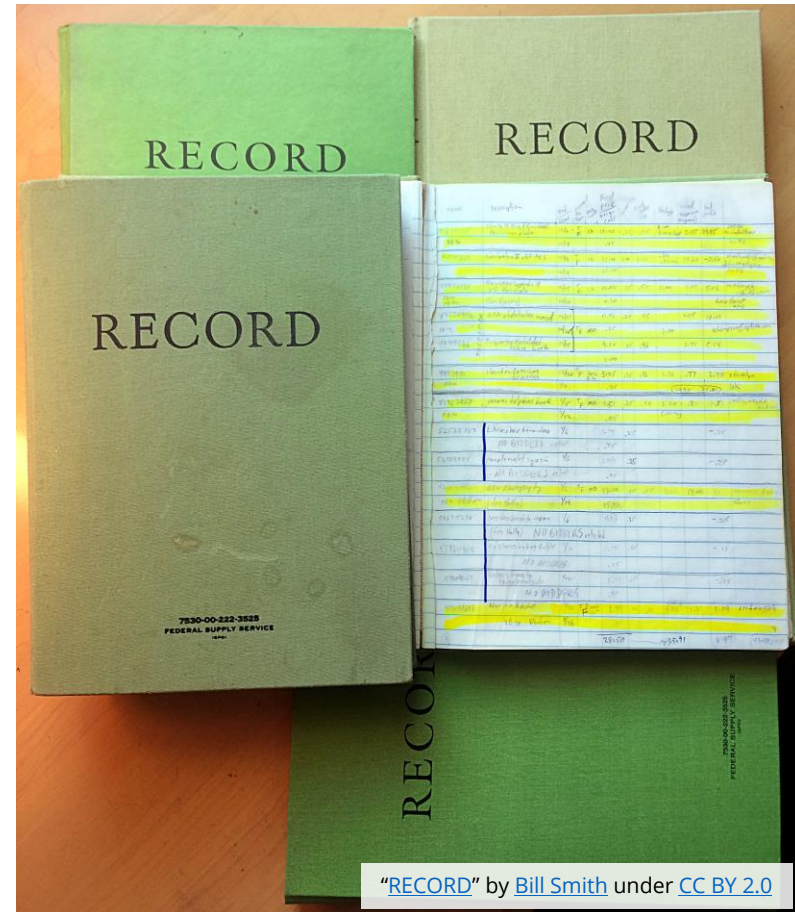
| Model   | Version   | Query type                | Date in response |
|---------|-----------|---------------------------|------------------|
| ChatGPT | GPT-4o    | Zero-shot                 | 2011-07-11       |
| ChatGPT | GPT-4o    | Screenshot interpretation | 2011-07-11       |
| Gemini  | 2.5 Flash | Zero-shot                 | can't say        |
| Gemini  | 2.5 Flash | Iterative                 | 2011-07-11       |
| Gemini  | 2.5 Flash | Screenshot interpretation | 2009 – 2012      |
| Grok    | 3         | Zero-shot                 | 2011-07-11       |
| Grok    | 3         | Screenshot interpretation | 2011-07-11       |
| Sonnet  | 4         | Zero-shot                 | 2011-07-11       |
| Sonnet  | 4         | Screenshot interpretation | 2011-07-11       |

# Roll-Up Observations

At least these model-versions preferred **the most obvious dates**; they seemed to ignore source code, URL date strings, IAWM capture dates, HTTP headers, etc.

They were most (unexpectedly) **useful in leveraging supplementary references.**

IAWM remains a critical source of ground-truth.



"RECORD" by Bill Smith under CC BY 2.0

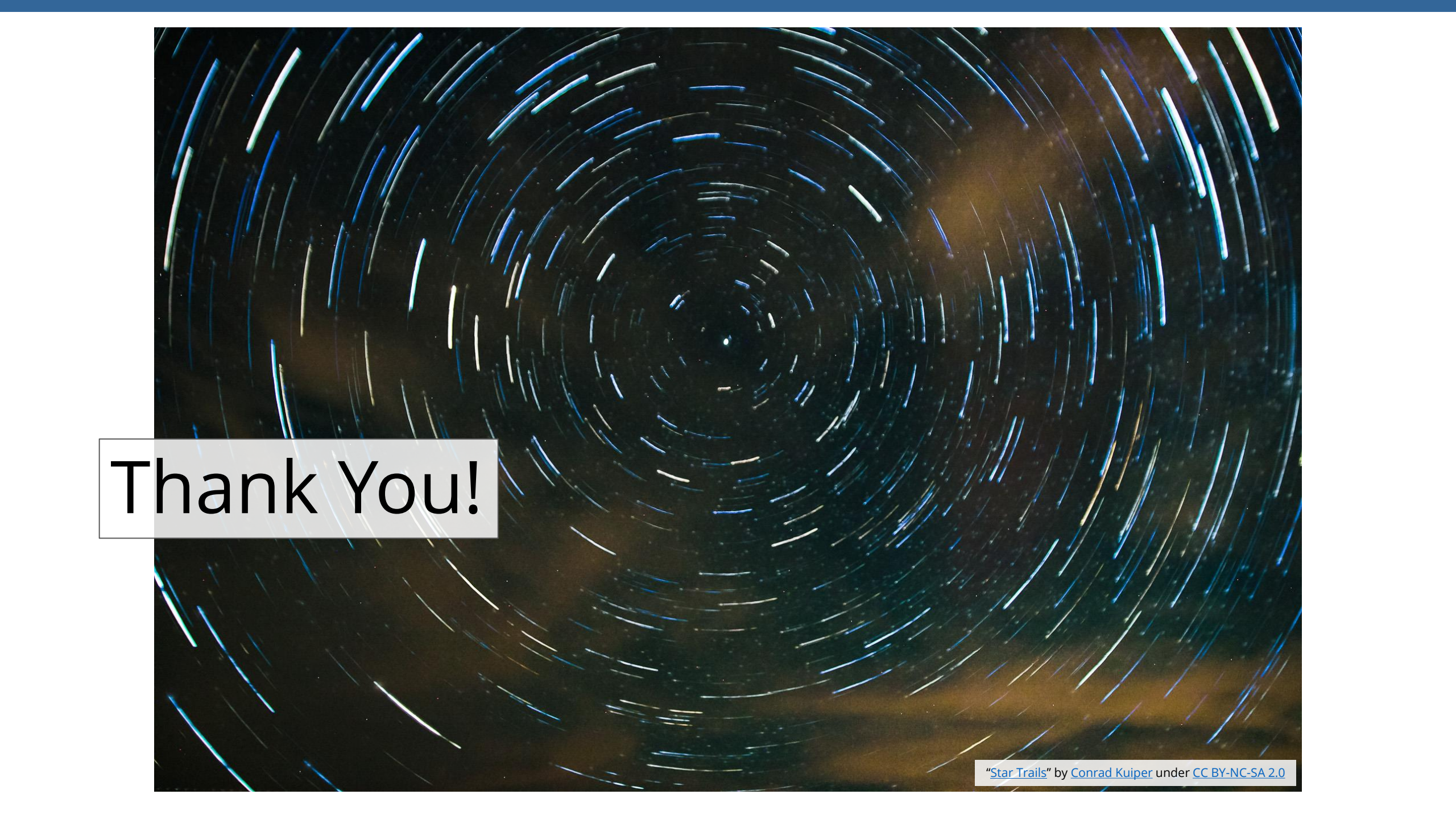
# Limitations

These models are several generations old.

I didn't test reasoning models.

I could've pushed them harder to evaluate other potential data signals.



A long-exposure photograph of the night sky showing concentric circular star trails. The trails are composed of many short, overlapping segments of light in shades of blue, cyan, and white, radiating from a central point. The background is a deep black with some faint, out-of-focus light sources.

# Thank You!