

Exploring Large Language Models for Analyzing Changes in Web Archive Content: A Retrieval-Augmented Generation Approach

Jhon G. Botello, Lesley Frew, Jose J. Padilla, Michele C. Weigle

Department of Computer Science
Virginia Modeling, Analysis, and Simulation Center
Old Dominion University, Norfolk, VA

December 17, 2024

✕ [@jhon_gbm12](#), [@lesley_elis](#), [@WebSciDL](#)

Both Researchers and the Public Are Interested in Tracking How Webpages Changes Over Time

The New York Times

Trump Transition | Updates | Cabinet Tracker | Tech Billionaires Steer Transition | Trump Defends Hegseth | Links Between Staff Picks

How Much Has ‘Climate Change’ Been Scrubbed From Federal Websites? A Lot.

Share full article

By **Coral Davenport**
Jan. 10, 2018

[nytimes.com](https://www.nytimes.com)

BuzzFeed News
REPORTING TO YOU

SIGN IN | ABOUT US | GOT A TIP? | BUZZFEED.COM

SCIENCE • CLIMATE CHANGE

Subscribe to BuzzFeed Daily Newsletter

Here’s How “Climate Change” Has Disappeared From Federal Websites Under Trump

Several agencies have scrubbed climate information from their websites, and often not communicated why, according to a new report. In many cases, it's unclear if the changes were mandated by Trump officials, self-censorship from career staff, or administrative tweaks without political motives.

Zahra Hirji
BuzzFeed News Reporter

Posted on January 10, 2018 at 11:51 am

[buzzfeednews.com](https://www.buzzfeednews.com)

ABOUT | TAKE ACTION | SCIENCE NETWORK | REPORTS & MULTIMEDIA | EN ESPAÑOL | DONATE

Union of Concerned Scientists | Climate | Energy | Transportation | Food | Nuclear Weapons | Science & Democracy

REPORTS & MULTIMEDIA / ATTACKS ON SCIENCE

Trump Administration Censored Information on Water Pollution, Climate Change, and Endangered Species

Published Jul 13, 2021

What happened: From 2017 to 2021, the Trump administration made approximately 1,400 changes to agency websites that removed science-based information on environmental issues, such as water pollution, climate

[ucsusa.org](https://www.ucsusa.org)



Storymodelers

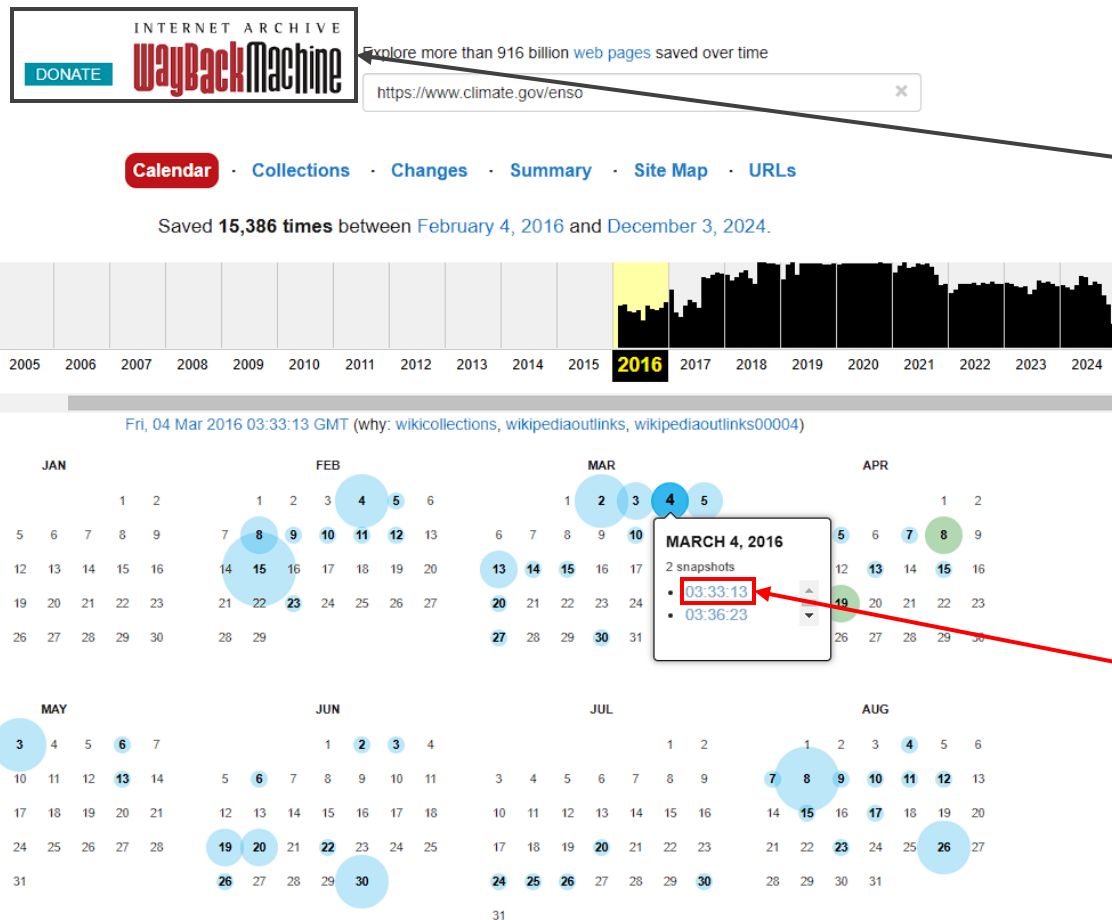


@WebSciDL

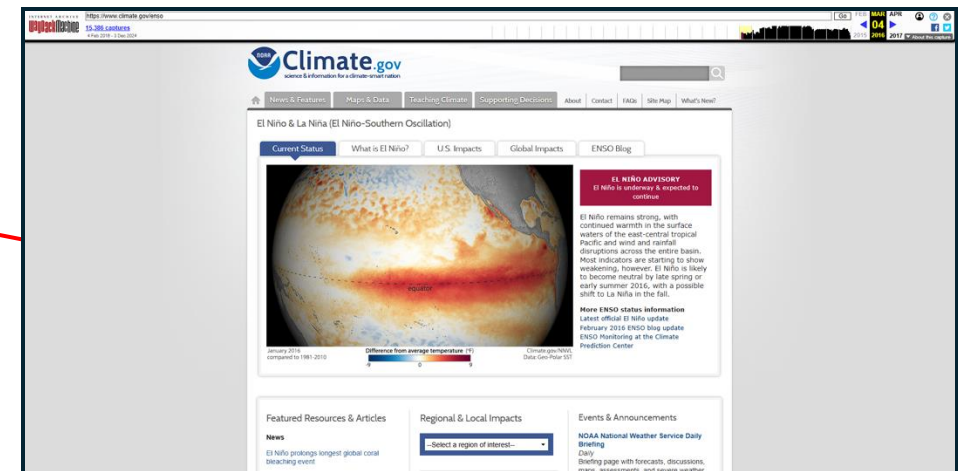


WebSciDL.bsky.social

Web Archives Preserve Historical Versions of Webpages, Capturing How They Appeared at a Specific Time



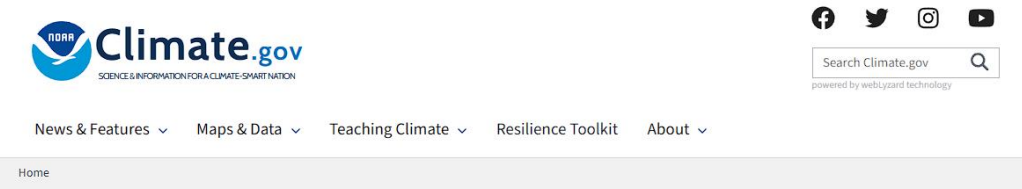
The Internet Archive's Wayback Machine is just one among many web archives.



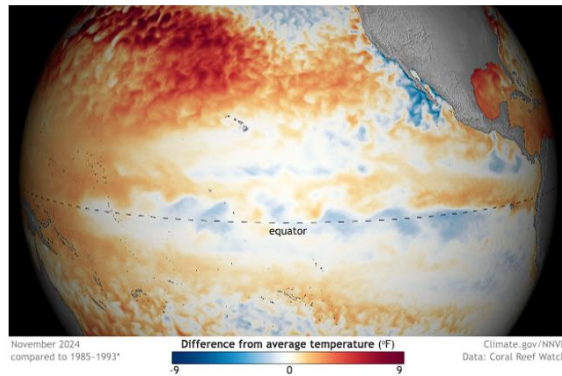
The Memento Protocol Connects the Live Web With the Archived Web

URI-R: Original resource on the live web
as of 2024-12-10

<https://www.climate.gov/enso>



El Niño & La Niña (El Niño-Southern Oscillation)



CURRENT STATUS | NOVEMBER 14, 2024

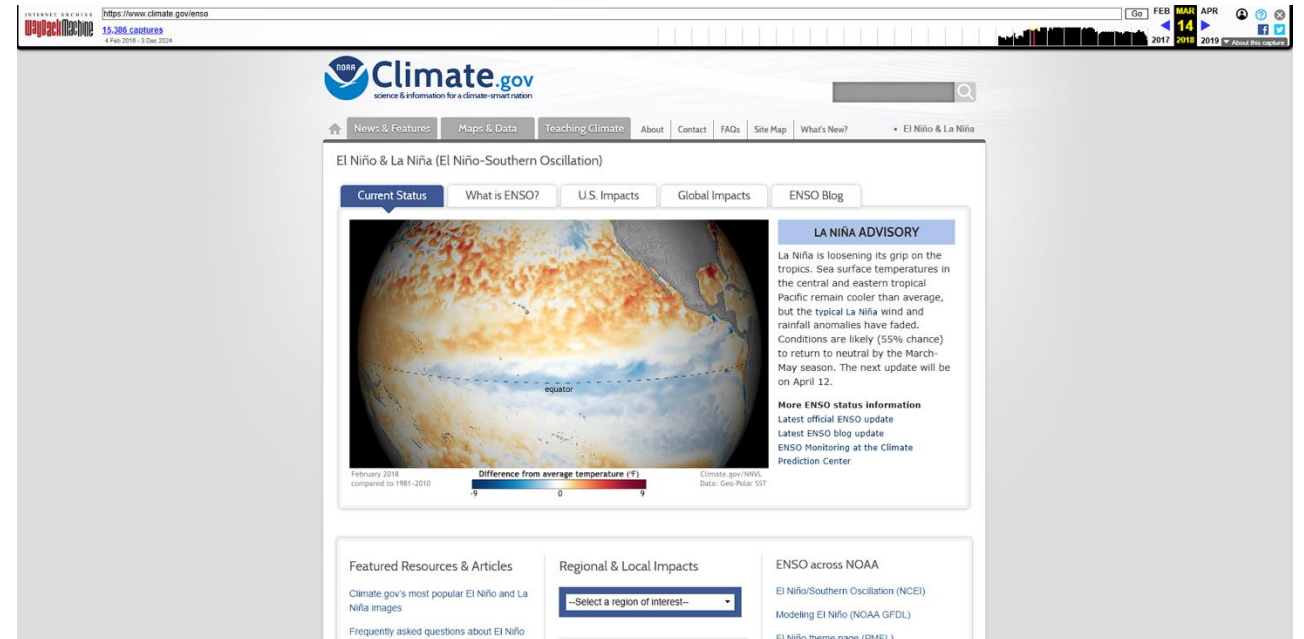
LA NIÑA WATCH

There's a 57% chance La Niña will develop soon. This is late for La Niña to arrive, and it's very likely to be a weak event at most. However, even a weak event can influence temperature, rain, and snow patterns across the world.

[Latest Official ENSO Update](#)
MORE ENSO

- [Latest ENSO blog](#)
- [ENSO in a nutshell](#)
- [ENSO FAQs](#)
- [Popular El Niño and La Niña images](#)

WHAT IS ENSO

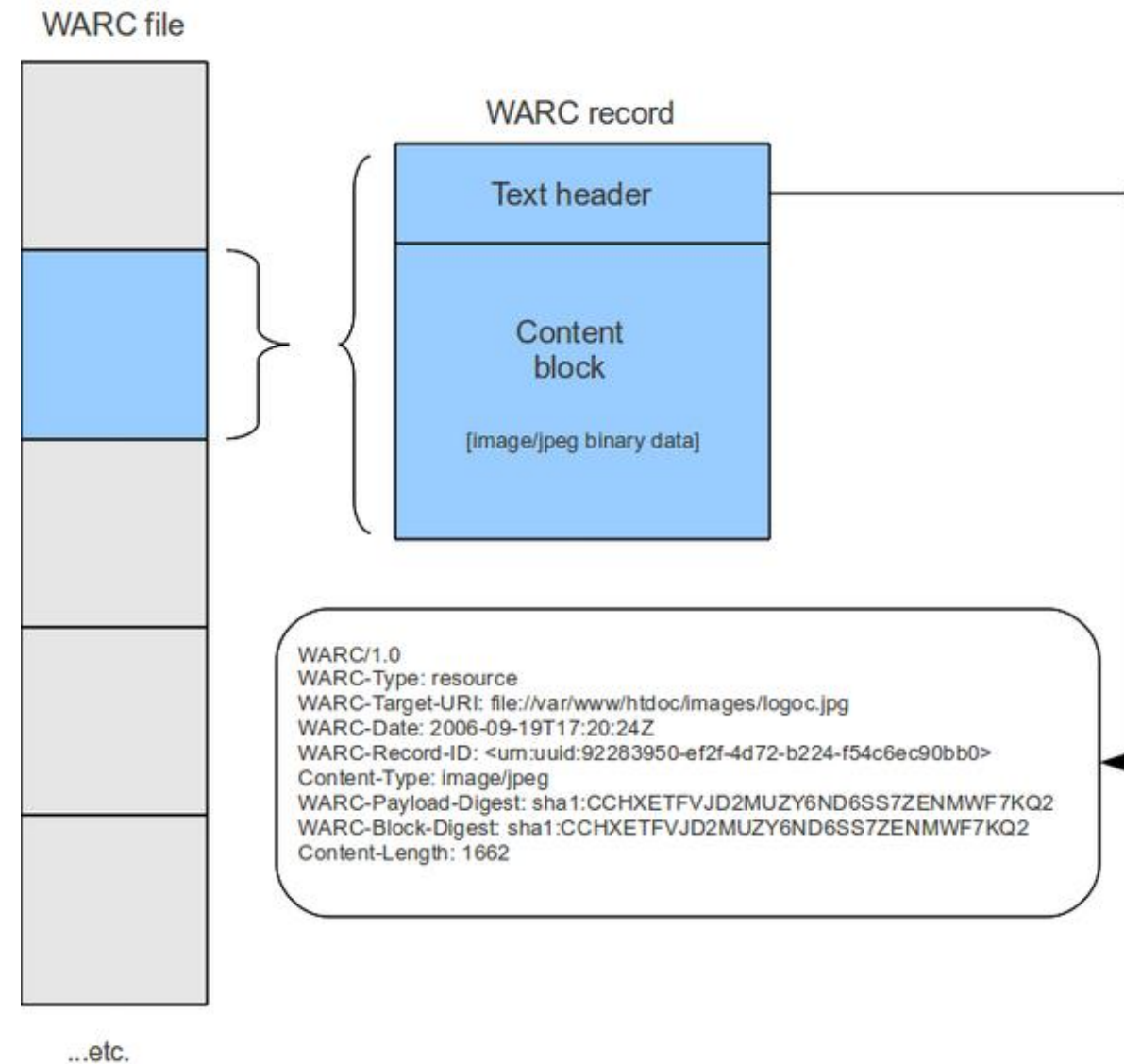


URI-M: Archived resource from 2018-03-14

<https://web.archive.org/web/20180314000741/climate.gov/enso>

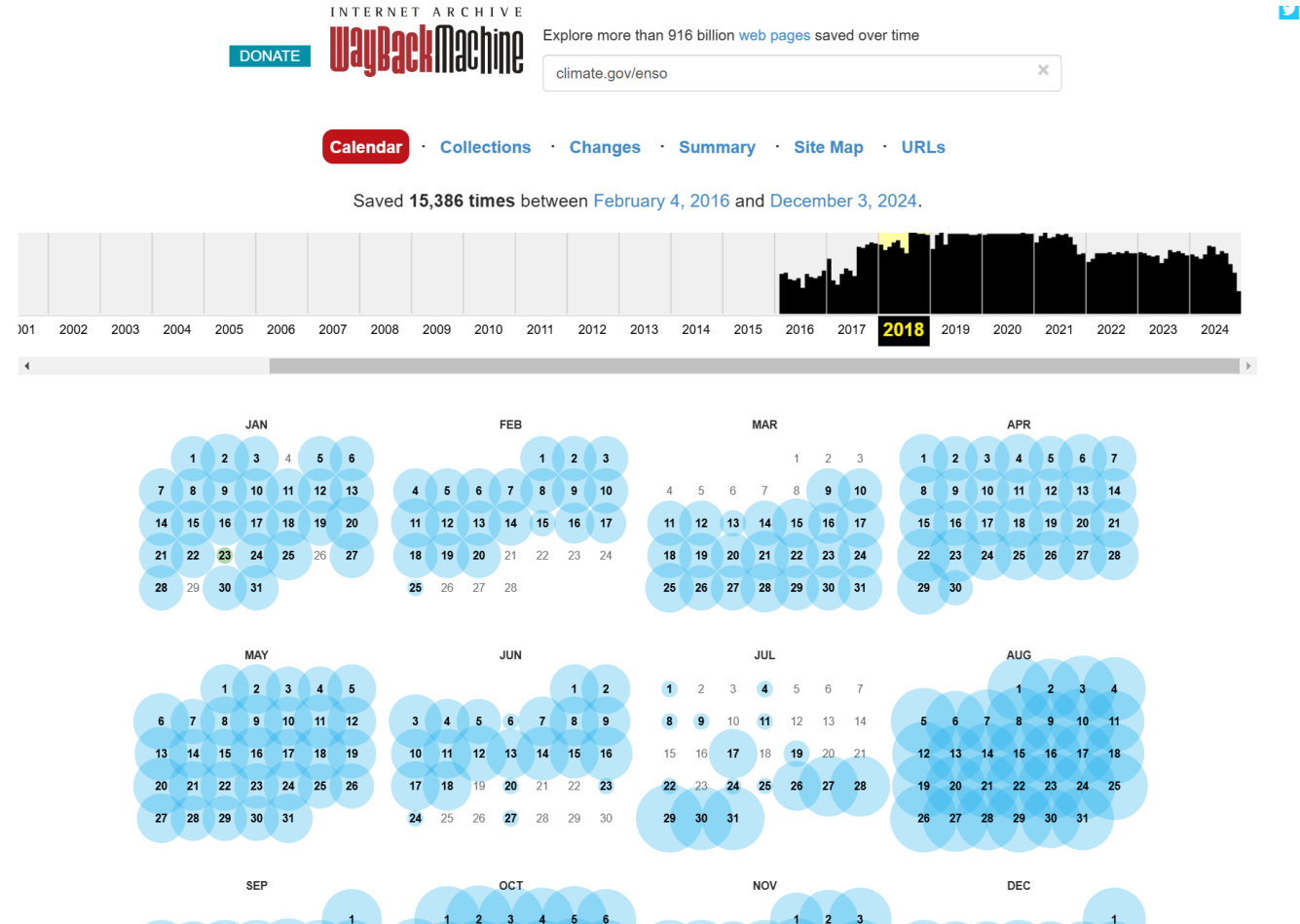
The Web ARChive (WARC) File Format Is Used To Store Mementos.

Each WARC file includes complete HTTP response headers, HTML content, and additional resources like images and JavaScript files

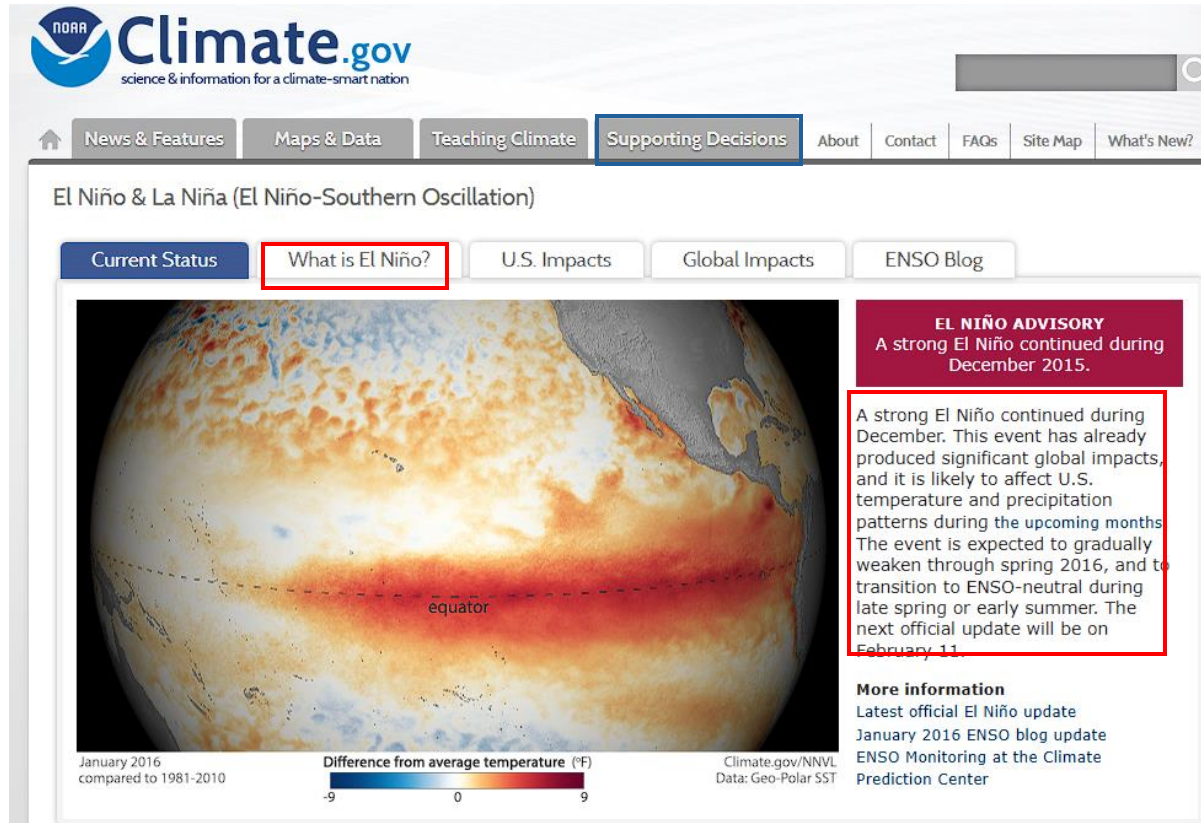


Source: https://wiki.archivematica.org/Significant_characteristics_of_websites

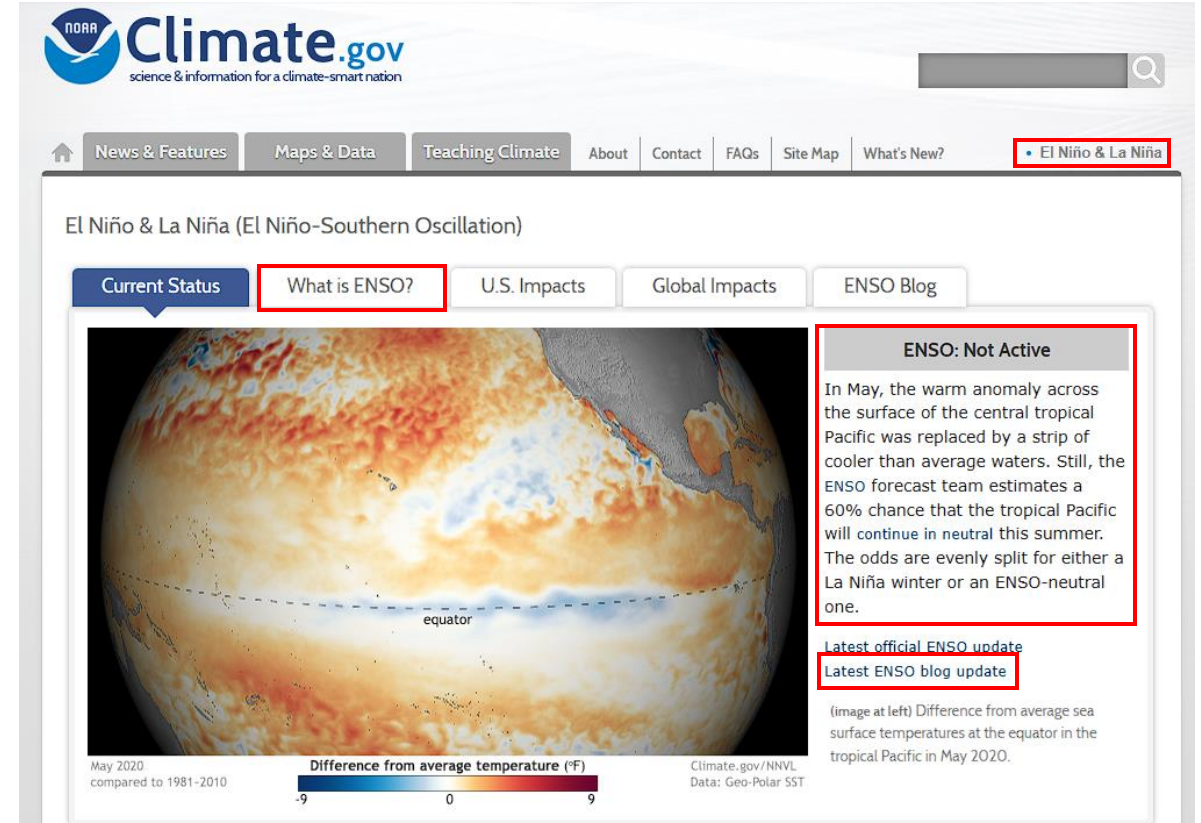
Finding Website Changes Is Often a Manual Process That Requires Significant Time and Effort



Analyzing Website Changes Is Often a Manual Process That Requires Significant Time and Effort



2016-02-04T07:23:26Z



2020-06-17T05:53:31Z

Target-URI: <https://www.climate.gov/enso>

Existing Tools Can't Be Used to Explain Complex Changes Like Content Structure Alterations or Semantic Context

- [Soy Infant Formula](#)
 - [NTP Evaluation](#)

- [Styrene](#)

- [Water Pollution](#)

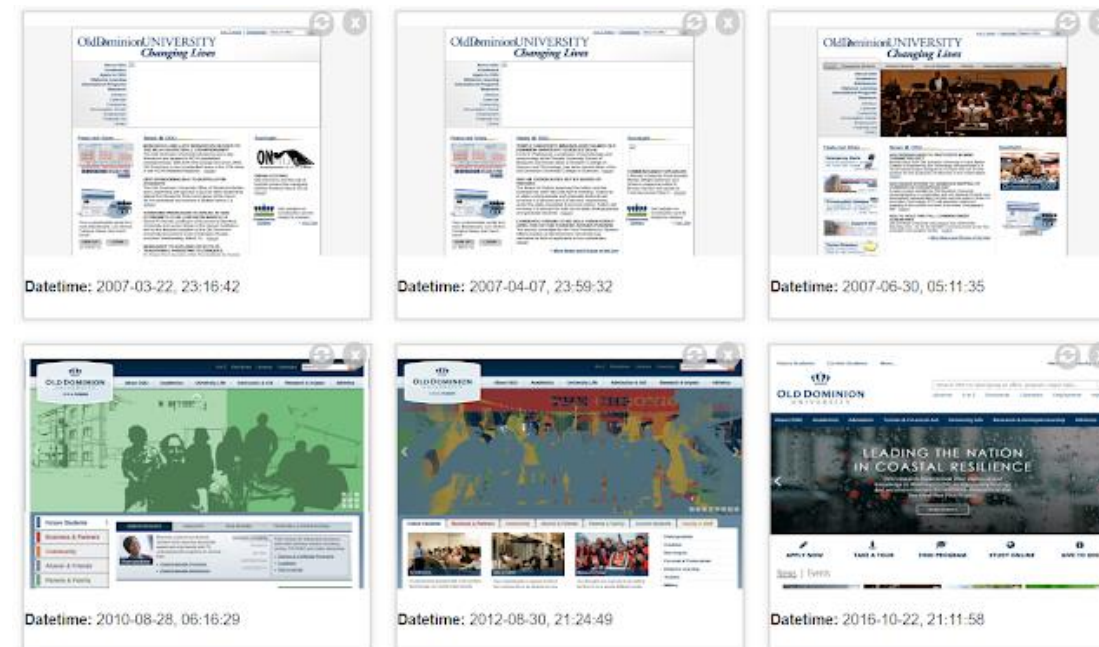
Environmental Health Links

- [The 13th Edition of the Report on Carcinogens](#) (324KB)
- [NTP Speaks About Aloe Vera](#) - The National Toxicology Program (NTP) conducted studies to help clarify the potential health hazards from ingestion of certain types of aloe vera.
- [Aloe Vera Fact Sheet](#) (1MB)
- [Concerned Citizens](#) - US EPA site geared towards citizens who want to become familiar with environmental issues and the potential environmental and human health risks caused by pollution. Covers important emergency phone numbers, health and safety issues at work, protecting children at home and a community's right to know about environmental exposures.
- [Environmental Defense Fund](#) - Environmental Defense Fund evaluates environmental problems and works to create and advocate solutions that win lasting political, economic and social support because they are nonpartisan, cost-efficient and fair. Topics include antibiotic resistance, agricultural policy, air quality, animal farms, environmental justice, pollution prevention, etc.

the most common form used in Aloe-based products.

- [Arsenic](#)
Arsenic is a naturally occurring element that is widely distributed in the Earth's crust. It is found in water, air, food, and soil.
- [Bisphenol A \(BPA\)](#)
An introduction to BPA and health | Bisphenol A (BPA) is a chemical produced in large quantities for use primarily in the production of polycarbonate plastics and epoxy resins
- [Cell Phone Radio Frequency Radiation](#)
The National Toxicology Program (NTP) headquartered at NIEHS is leading the largest laboratory rodent study to date on cell phone radio frequency. NTP studies will help clarify any potential health hazards from exposure to cell phone radiation.
- [Climate Change](#)
Climate change is the result of the buildup of greenhouse gases in the atmosphere, primarily from the burning of fossil fuels for energy and other human activities. These gases, such as carbon dioxide and methane, warm and alter the global climate, which causes environmental changes to occur that can harm people's health and well-being.

The [Wayback Machine Changes Tool](#) allows users to view term based differences between webpages at known datetimes



TMVis allows users to view visual differences between webpages over time

Mabe, A et al. (2020). TMVis: Visualizing Webpage Changes Over Time.

Manually Examining Changes to this Webpage will be Tedious and Difficult

INTERNET ARCHIVE <https://earthdata.nasa.gov/user-resources/remote-sensors> Go MAR APR NOV 16 2016 2017 About this capture

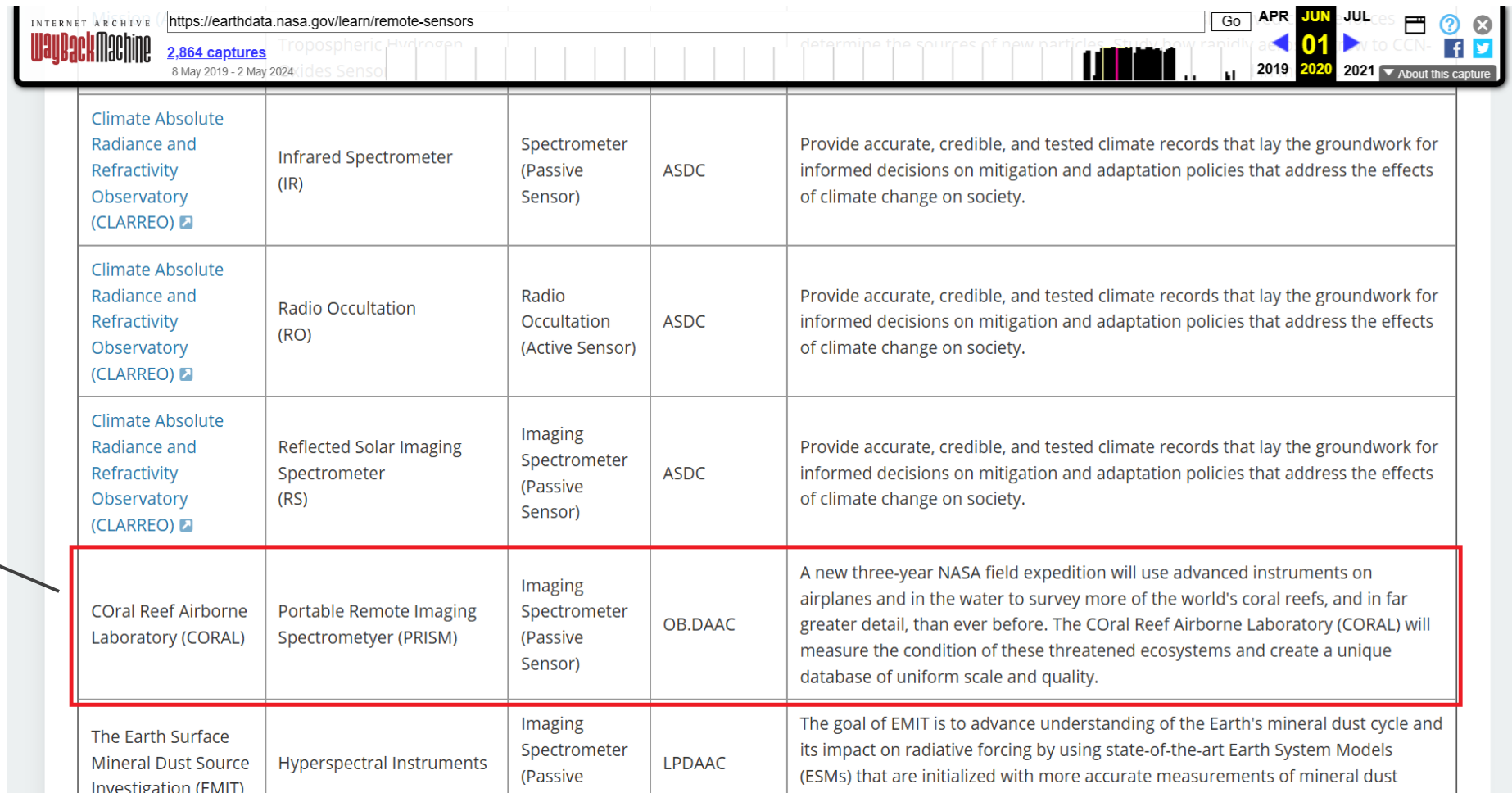
WayBackMachine 3,579 captures 17 Jul 2015 - 9 Oct 2024

Feedback




Climate Absolute Radiance and Refractivity Observatory (CLARREO)	Infrared Spectrometer (IR)	Spectrometer (Passive Sensor)	ASDC DAAC	Provide accurate, credible, and tested climate records that lay the groundwork for informed decisions on mitigation and adaptation policies that address the effects of climate change on society.
Climate Absolute Radiance and Refractivity Observatory (CLARREO)	Radio Occultation (RO)	Radio Occultation (Active Sensor)	ASDC DAAC	Provide accurate, credible, and tested climate records that lay the groundwork for informed decisions on mitigation and adaptation policies that address the effects of climate change on society.
Climate Absolute Radiance and Refractivity Observatory (CLARREO)	Reflected Solar Imaging Spectrometer (RS)	Imaging Spectrometer (Passive Sensor)	ASDC DAAC	Provide accurate, credible, and tested climate records that lay the groundwork for informed decisions on mitigation and adaptation policies that address the effects of climate change on society.
Cyclone Global Navigation Satellite System (CYGNSS)	Delay Doppler Mapping Instrument (DDMI)	Scatterometer (Active Sensor)	PO.DAAC	Measure Ocean surface wind speed in all precipitating conditions, including those experienced in the tropical Cyclone (TC) eyewall. Measure ocean surface wind speed in the TC inner core with sufficient frequency to resolve genesis and rapid intensification processes. This study will focus to understand the coupling between ocean surface properties, moist atmospheric thermodynamics, radiation, and

Users Currently Must Manually Check if Changes Are Semantically Meaningful, but LLMs Offer a Novel Approach

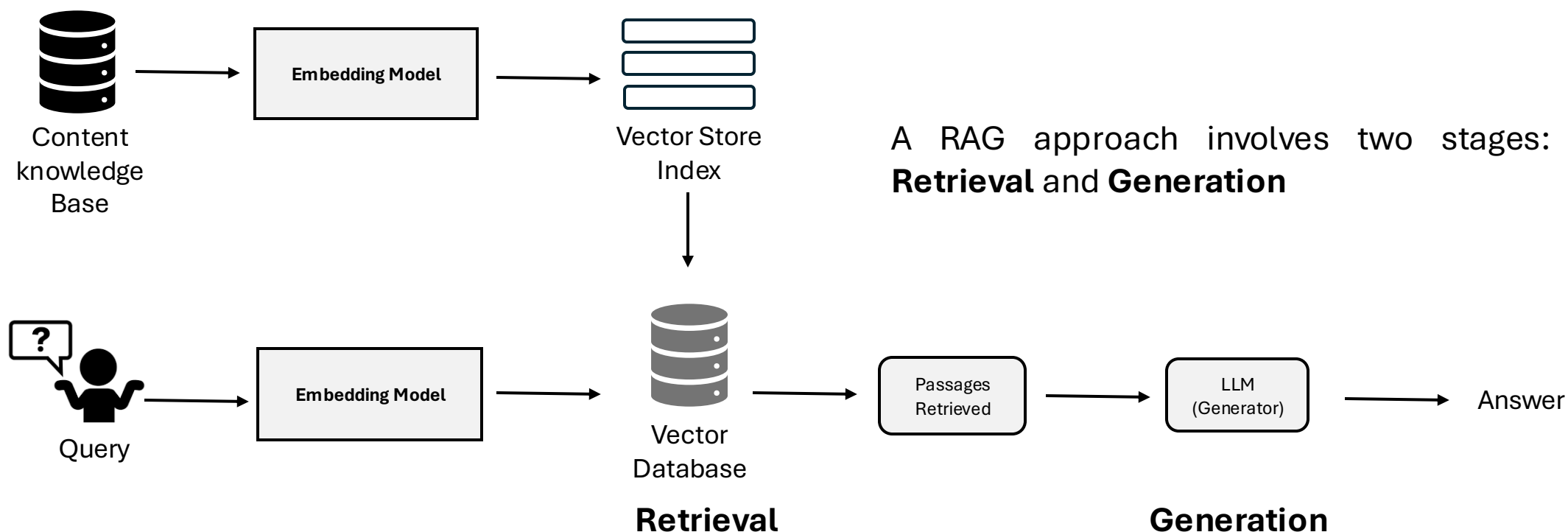
Researchers can use LLMs to detect semantic shifts: in this case, from climate modeling to broader ecological monitoring, like coral reef health tracking.



Wayback Machine 2,864 captures 8 May 2019 - 2 May 2024

Climate Absolute Radiance and Refractivity Observatory (CLARREO) 	Infrared Spectrometer (IR)	Spectrometer (Passive Sensor)	ASDC	Provide accurate, credible, and tested climate records that lay the groundwork for informed decisions on mitigation and adaptation policies that address the effects of climate change on society.
Climate Absolute Radiance and Refractivity Observatory (CLARREO) 	Radio Occultation (RO)	Radio Occultation (Active Sensor)	ASDC	Provide accurate, credible, and tested climate records that lay the groundwork for informed decisions on mitigation and adaptation policies that address the effects of climate change on society.
Climate Absolute Radiance and Refractivity Observatory (CLARREO) 	Reflected Solar Imaging Spectrometer (RS)	Imaging Spectrometer (Passive Sensor)	ASDC	Provide accurate, credible, and tested climate records that lay the groundwork for informed decisions on mitigation and adaptation policies that address the effects of climate change on society.
Coral Reef Airborne Laboratory (CORAL)	Portable Remote Imaging Spectrometryer (PRISM)	Imaging Spectrometer (Passive Sensor)	OB.DAAC	A new three-year NASA field expedition will use advanced instruments on airplanes and in the water to survey more of the world's coral reefs, and in far greater detail, than ever before. The CORal Reef Airborne Laboratory (CORAL) will measure the condition of these threatened ecosystems and create a unique database of uniform scale and quality.
The Earth Surface Mineral Dust Source Investigation (EMIT)	Hyperspectral Instruments	Imaging Spectrometer (Passive)	LPDAAC	The goal of EMIT is to advance understanding of the Earth's mineral dust cycle and its impact on radiative forcing by using state-of-the-art Earth System Models (ESMs) that are initialized with more accurate measurements of mineral dust

Using LLMs for Change Analysis Is Promising, With Archived Web Content Changes Still Largely Unexplored



Why Is This Approach Helpful?

Traditional Methods

- ❖ Token to token text comparison
- ❖ Pairwise comparison of mementos
- ❖ Specialized algorithms/Tools
- ❖ Keyword searches and metadata filters
- ❖ Structural approach

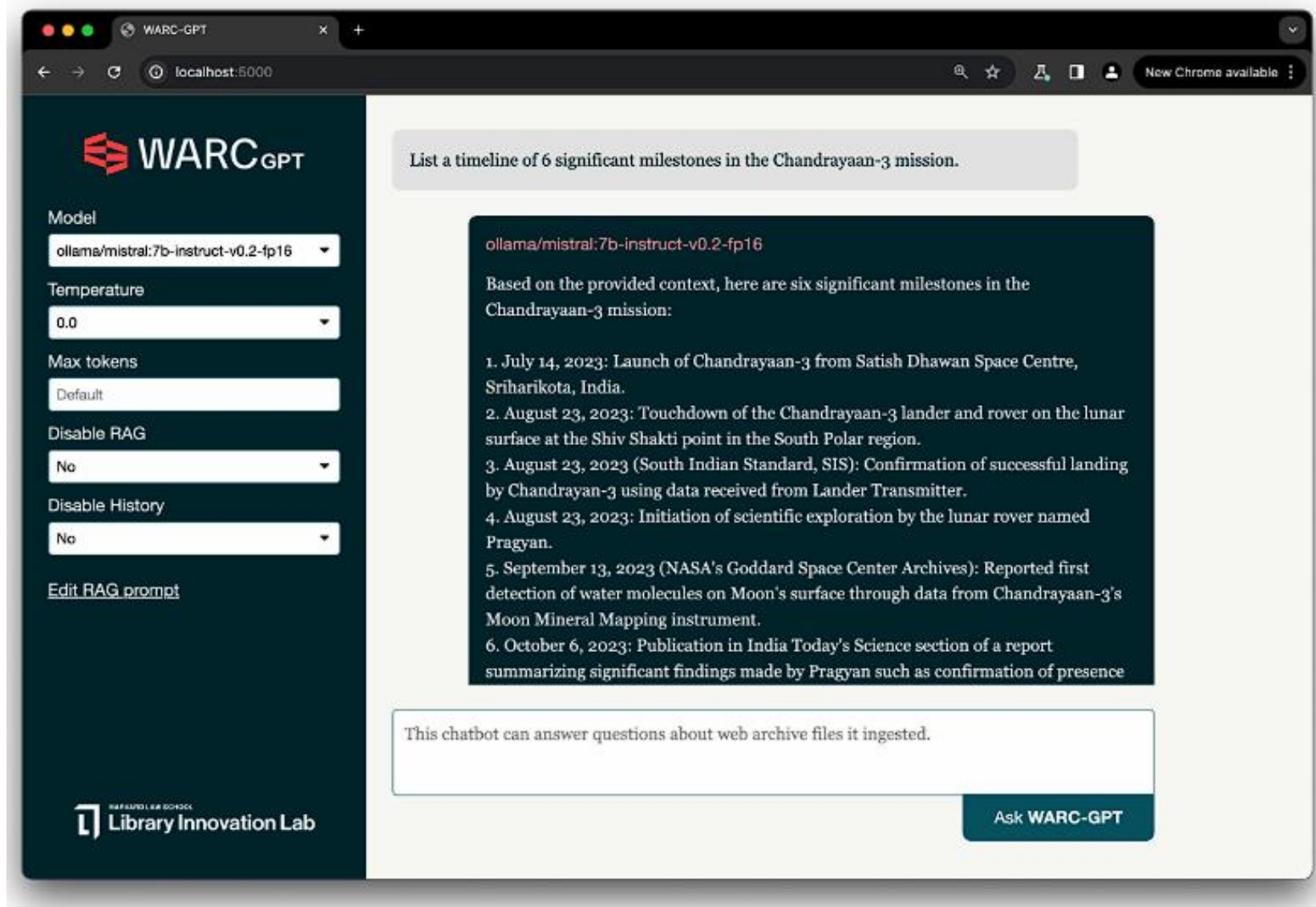
RAG

- ❖ Analysis based on natural language
- ❖ Comparison among multiple mementos
- ❖ Q&A Interaction
- ❖ Does not depend on a term/keyword
- ❖ Semantic approach

WARC-GPT's RAG Approach Enables Creating Custom Chatbots That Use Web Archive Collections as a Knowledge Base

...but can it be used to analyze changes?

Harvard Law School Library Innovation Lab. (2024, February 12). **WARC-GPT: An open-source tool for exploring web archives with AI.** Harvard Law School Library Innovation Lab Blog. <https://lil.law.harvard.edu/blog/2024/02/12/warc-gpt-an-open-source-tool-for-exploring-web-archives-with-ai/>



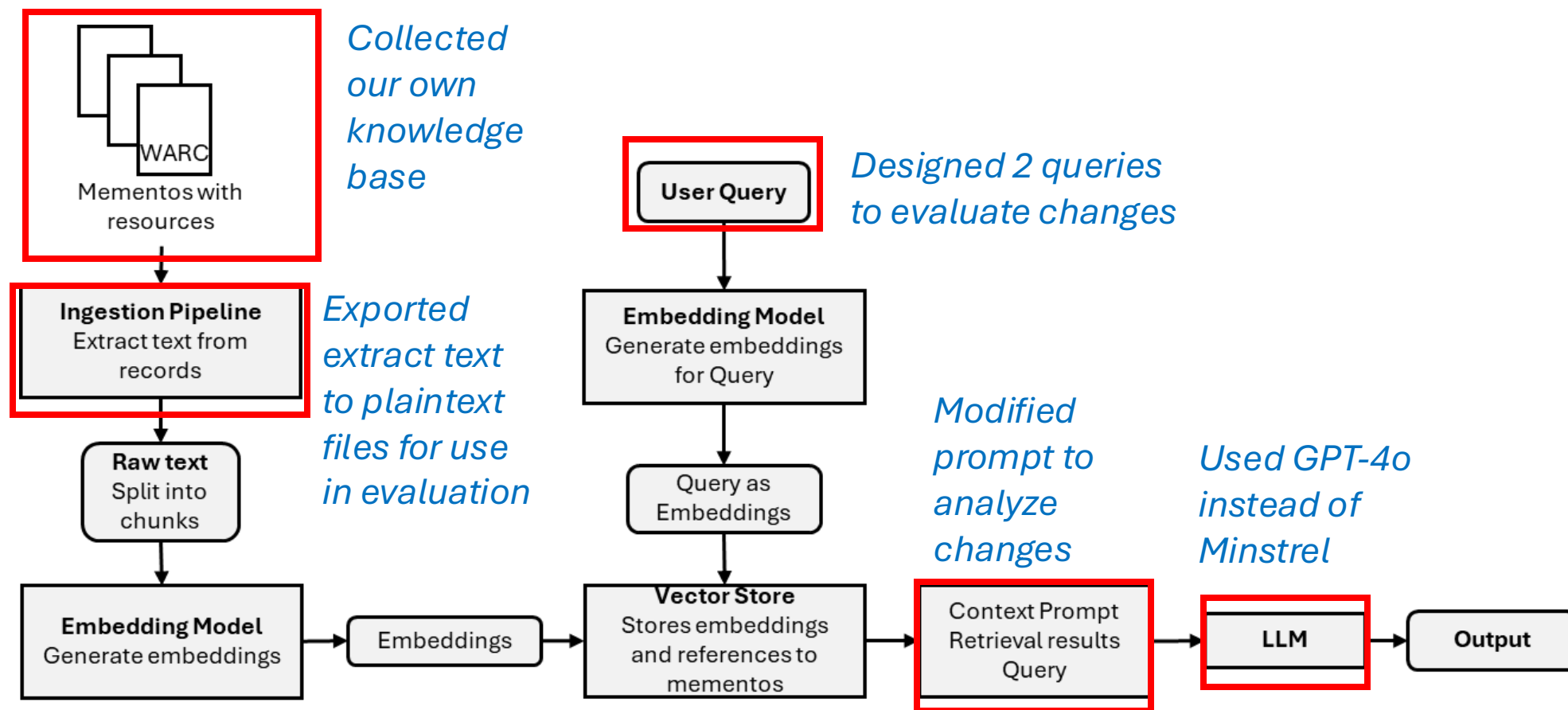
We Investigated 3 Research Questions:

How can the use of LLMs make changes in webpages discoverable and understandable?

What advantages/limitations does a RAG-based approach have compared to traditional methods?

To what extent are researchers able to use output from WARC-GPT with GPT-4o to identify changes/inconsistencies?

We Explored the System's Ability To Retrieve Changed Passages and Assessed GPT-4o's Output



WARC-GPT Architecture Considering a Set of Mementos as a Knowledge Source.

<https://lil.law.harvard.edu/blog/2024/02/12/warc-gpt-an-open-source-tool-for-exploring-web-archives-with-ai/>

We Applied Prompt Engineering To Adjust and Refine the Context Prompt per Default, Ensuring It Contextualizes the Task Effectively

Modified Context Prompt

How can you help me?

I can help you by analyzing the changes across different archived versions of a webpage. By comparing these versions, I can identify content changes such as alterations in terms, phrases, or paragraphs, as well as changes in the semantic Context. This analysis can provide insights into how the webpage's content and focus have evolved over time.

```
TEXT_COMPLETION_BASE_PROMPT = "  
{history}
```

```
You are a helpful assistant that analyzes changes among archived versions of webpages (Mementos).  
Your task is to compare different versions of webpages, identify changes, and provide a clear analysis of the differences.
```

```
{rag}
```

```
Request: {request}
```

```
Helpful response (plain text, no markdown):  
"
```

```
TEXT_COMPLETION_RAG_PROMPT = "
```

```
Here is context to help you fulfill the user's request:  
{context}
```

```
-----  
The context comes from different Mementos (archived versions) of the same webpage.  
Use the context to identify and explain changes across the different versions.  
When analyzing changes, focus on: 1) Content changes such as alterations in terms, phrases, or paragraphs, and 2) Changes  
in the semantic context.  
If the context lacks sufficient information to identify changes, simply state that there is insufficient information without  
speculating or making assumptions.  
Ignore irrelevant or empty context.  
Cite and quote your sources whenever possible. Use their number (for example: [1]) and / or URL to reference them.  
"
```

Our Knowledge Base Contained Mementos From U.S. Federal Environmental Webpages That Changed Between 2016 and 2020



10 Websites



8 Websites



U.S. DEPARTMENT OF
ENERGY

3 Websites



EPA United States
Environmental Protection
Agency

9 Websites

We Empirically Designed Two Queries To Analyze How Well The System Could Be Used To Detect Changes and Provide Analysis.



Query 1

What Are the Changes or Inconsistencies in Terms, Phrases, or Paragraphs Among the Mementos of Webpage X?



Query 2

Compare Any Change Regarding the Semantic Context Among the Mementos for the Webpage X.

Answers were validated through human evaluation, ensuring their accuracy about the corresponding elements across mementos.

Answers available at <https://github.com/jgbotello/Exploring-LLMs-for-Content-Change-Analysis>

Answer To Query 1 Details Content and Structure Changes, While Answers to Query 2 Focuses on Their Semantic Significance

Agency	URI-R	Mementos	Answer
NASA	https://airbornescience.nasa.gov/	2	...In the 2016 version, the webpage includes sections such as "Program Current Activities," "Student Airborne Research Program," "ASP Blog," and "Research Opportunities" [1]. By 2020, the structure has shifted to include "Airborne Missions," "NASA Earth Expeditions Blog," and "ASP Sensor Network" [4]. This indicates a reorganization and possible expansion of the program's focus areas...
DOE	http://srnl.doe.gov/	2	...In the 2016 version, there is a mention of a partnership with Clemson to improve the electric grid. The 2020 version highlights SRNL seeking partners for the DOE Visiting Faculty Program....
EPA	https://www.epa.gov/climate-change-water-sector	4	...The 2016 versions (both January and June) include a "Climate Change and Water Newsletter" section, which is absent in the 2020 version, previously offering newsletter information...

Example of Answer to Query 1

Agency	URI-R	Mementos	Answer
NASA	https://airbornescience.nasa.gov/	2	...The changes from 2016 to 2020 reflect a broadening of the information available on the site, with a greater emphasis on operational details and user engagement, while maintaining a strong focus on research and educational opportunities...
DOE	http://srnl.doe.gov/	2	...Changes reflects a move from highlighting specific events and partnerships to emphasizing ongoing research and development initiatives and broader programmatic goals...
EPA	https://www.epa.gov/climate-change-water-sector	4	...These changes in the semantic context reflect updates in the organization, focus, and presentation of information on the webpage over time...

Example of Answer to Query 2

The Model Could Be Used to Identify Changes in Content and Structure

- 1 page had a generated answer that did not mention any changes to content
- 16 pages had a generated answer that did not mention any changes to structure
- 1 page had a generated answer that included changes not on the page (they were changes for WP22)

Webpage (WP)	Content	Structure	Accurate?
WP1	1	0	Y
WP2	1	1	Y
WP3	1	0	Y
WP4	1	0	Y
WP5	1	1	Y
WP6	1	0	Y
WP7	1	1	Y
WP8	1	0	Y
WP9	1	1	Y
WP10	1	1	Y
WP11	0	1	Y
WP12	1	0	Y
WP13	1	0	Y
WP14	1	0	Y
WP15	1	0	Y
WP16	1	1	Y
WP17	1	0	Y
WP18	1	0	Y
WP19	1	0	Y
WP20	1	1	Y
WP21	1	0	N
WP22	1	0	Y
WP23	1	0	Y
WP24	1	0	Y

Summary of changes detected across webpages

This Exploratory Study is Limited by Dataset Size and Computational Resources



Dataset Size Limited to 30 web pages and 67 WARC files



Embedding generation models and LLMs require substantial computational power, memory, and cost



When pages belong to the same domain, the system may extract incorrect passages, necessitating manual double-checking.

Future Work will Incorporate Non-Textual Changes and Evaluate The Use of Alternative Large Language Models

Explore the integration of multimodal models to analyze text and images in WARC files.

Investigate the use of various LLMs, such as ClimateGPT and LLM360's Open-Source K2, to evaluate differences in responses and their impact on web archiving research.

Exploring Large Language Models for Analyzing Changes in Web Archive Content: A Retrieval-Augmented Generation Approach



Explored the use of GPT-4o with WARC-GPT, a Retrieval-Augmented Generation pipeline, for detecting and analyzing changes in archived webpages.



Demonstrated that LLMs enable deeper analysis by capturing semantic context, unlike traditional methods focused on term frequency or surface-level differences.



Conducted a case study on U.S. Federal Agency webpages related to environmental issues, analyzing changes between 2016 and 2020.



Acknowledged limitations such as computational demands and challenges in retrieving relevant passages.



Established a foundation for future research in webpage change analysis, advancing the integration of web archiving and AI.

<https://oduwsdl.github.io/>

Backup slides

The Default Prompt Does Not Effectively Allow for the Analysis of the Changes Between Multiple Captures of a Webpage

WARC-GPT Context Prompt per Default

How can you help me?

I can assist you by providing information, answering questions, and offering guidance on a wide range of topics. If you have specific questions or need help with something, feel free to ask!

```
TEXT_COMPLETION_BASE_PROMPT = "  
{history}  
  
You are a helpful assistant.  
  
{rag}  
  
Request: {request}  
  
Helpful response (plain text, no markdown):  
"  
  
# NOTE: Injected into BASE prompt when relevant.  
# Inspired by LangChain's default RAG prompt.  
# {context} is a reserved keyword.  
TEXT_COMPLETION_RAG_PROMPT = "  
Here is context to help you fulfill the user's request:  
{context}  
-----  
Context comes from web pages that were captured as part of a web archives collection.  
When possible, use context to answer the question asked by the user.  
If you don't know the answer, just say that you don't know, don't try to make up an answer.  
Ignore context if it is empty or irrelevant.  
Cite and quote your sources whenever possible. Use their number (for example: [1]) and / or URL to reference them.  
"
```

Of the 30 Webpages Examined, the System Retrieved 24 Relevant Passages Using Query 1

What Are the Changes or Inconsistencies in Terms, Phrases, or Paragraphs Among the Mementos of Webpage X?

Category	Count
Total pages analyzed	30
Pages with Response	
Pages with answer to Query 1	24
Pages without answer to Query 1	
With generated embeddings	5
Embedding generation omitted	1

Summary of Query 1 Response Analysis

AI

There are several changes and inconsistencies among the different mementos of the webpage: 1. ****ENSO Status****: - The earlier memento from June 2020 mentions that ENSO is "NOT ACTIVE" with a 60% chance of remaining neutral throughout the summer and evenly split odds for a La Niña or ENSO-neutral winter. - The later memento from December 2020 provides a "La Niña Advisory" indicating that La Niña strengthened in the tropical Pacific in October 2020, with a 95% chance it will last through the Northern Hemisphere winter. 2. ****Temperature Conditions****: - The June 2020 memento describes a strip of cooler-than-average waters replacing the earlier warm anomaly in the central tropical Pacific. - In the December 2020 memento, cooler-than-average surface waters are mentioned as a sign of La Niña. 3. ****Impacts on the United States****: - The December 2020 memento notes that La Niña winters tend to bring warm and dry conditions to the southern tier of the U.S. and snowier-than-average conditions to the northern U.S. These changes reflect a transition from an ENSO-neutral phase in mid-2020 to a La Niña phase by the end of 2020, along with the associated differences in weather impacts and sea surface conditions.

[SOURCE](#)

Example of Answer

Examples of Excluded Pages

5 pages with
“insufficient
information”
response per
prompt

11. URI-R: <https://www3.epa.gov/ttn/chief/comments.html> [2 mementos]

Answer Query 1: There is insufficient information provided in the context to identify changes regarding the semantic context among the mementos of the webpage www3.epa.gov/ttn/chief/comments.html. The excerpts provided do not include content from this specific webpage, so no analysis can be conducted.

Answer Query 2: There is insufficient information provided in the context to identify changes regarding the semantic context among the mementos of the webpage www3.epa.gov/ttn/chief/comments.html. The excerpts provided do not include content from this specific webpage, so no analysis can be conducted.

1 page with
an exception
response

15. URI-R: <https://earthdata.nasa.gov/user-resources/remote-sensors> [3 mementos]

Answer Query 1: The context provided does not include multiple mementos or versions of the webpage "earthdata.nasa.gov/user-resources/remote-sensors" to compare. Therefore, there is insufficient information to identify changes or inconsistencies in terms, phrases, or paragraphs among the mementos of this specific webpage.

Answer Query 2: The archived versions of the webpage <https://earthdata.nasa.gov/user-resources/remote-sensors> show several changes in content and semantic context over time:

1. Content Changes:

- The 2016-03-12 version [1] includes detailed descriptions of various climate observation instruments and their purposes, such as the Climate Absolute Radiance and Refractivity Observatory (CLARREO) and the Cyclone Global Navigation Satellite System (CYGNSS). It focuses on providing accurate climate records and understanding climate change impacts.

The Model Could Be Used To Identify Changes in Content And Content Structure, Even When Considering More Than 2 Mementos

Agency	URI-R	Mementos	Answer
NASA	https://airbornescience.nasa.gov/	2	...In the 2016 version, the webpage includes sections such as "Program Current Activities," "Student Airborne Research Program," "ASP Blog," and "Research Opportunities" [1]. By 2020, the structure has shifted to include "Airborne Missions," "NASA Earth Expeditions Blog," and "ASP Sensor Network" [4]. This indicates a reorganization and possible expansion of the program's focus areas...
DOE	http://srnl.doe.gov/	2	...In the 2016 version, there is a mention of a partnership with Clemson to improve the electric grid. The 2020 version highlights SRNL seeking partners for the DOE Visiting Faculty Program....
EPA	https://www.epa.gov/climate-change-water-sector	4	...The 2016 versions (both January and June) include a "Climate Change and Water Newsletter" section, which is absent in the 2020 version, previously offering newsletter information...

We Included more than 2 mementos for some webpages

Example of Answer to Query 1

Answer To Query 1 Details Content and Structure Changes, While Answers to Query 2 Focuses on Their Semantic Significance

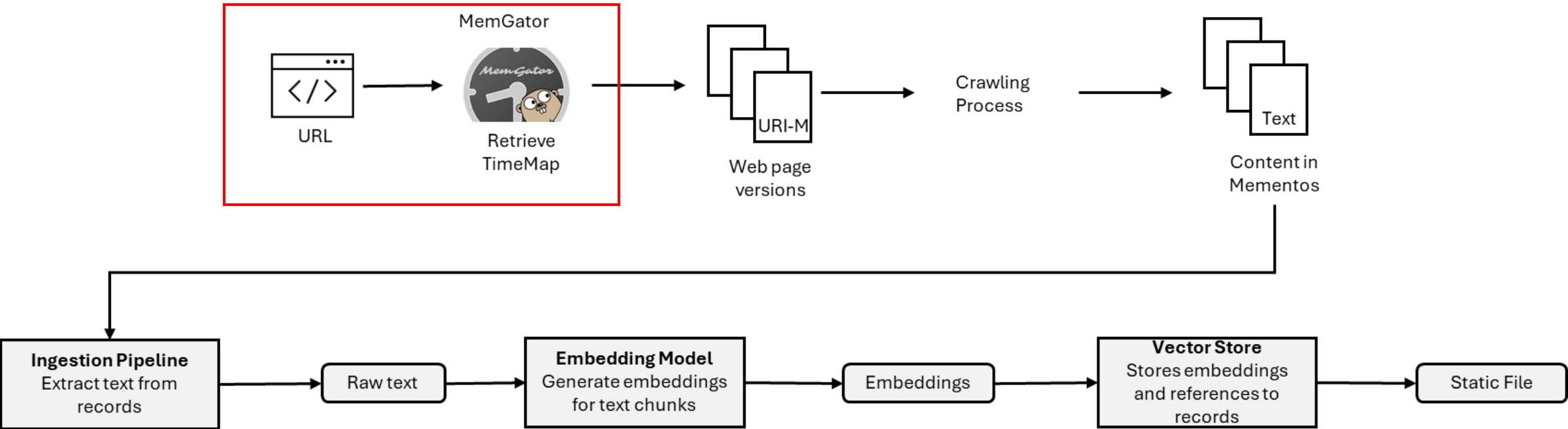
Agency	URI-R	Mementos	Answer
NASA	https://airbornescience.nasa.gov/	2	...The changes from 2016 to 2020 reflect a broadening of the information available on the site, with a greater emphasis on operational details and user engagement, while maintaining a strong focus on research and educational opportunities...
DOE	http://srnl.doe.gov/	2	...Changes reflects a move from highlighting specific events and partnerships to emphasizing ongoing research and development initiatives and broader programmatic goals...
EPA	https://www.epa.gov/climate-change-water-sector	4	...These changes in the semantic context reflect updates in the organization, focus, and presentation of information on the webpage over time...

Example of Answer to Query 2

We should interpret LLM responses with caution

The semantic context of the webpage www3.epa.gov/ttn/atw/ has undergone a change between the archived versions from 2016 and 2020. In the 2016 version, the section “rules & implementation” is referred to as “national- scale air toxics assessment” [3]. By 2020, this section is referred to as “national air toxics assessment” [2]. This change in terminology from “national- scale” to “national” suggests a possible shift in focus or scope, potentially indicating a broader or more inclusive approach to air toxics assessment. However, without additional context, the exact implications of this change in terminology are not fully clear.

Future Work



Modified WARC-GPT Ingestion Pipeline

The modified WARC GPT ingestion pipeline will eliminate the need for users to collect WARCS and allow for URL based input instead