

★ Get unlimited access to the best of Medium for less than \$1/week. [Become a member](#)



In-Text Citing for RAG Question-Answering



Yotam Abraham · [Follow](#)

5 min read · Jul 6, 2023



Listen



Share



More



<https://flickr.com/photos/87913776@N00/5129607997>

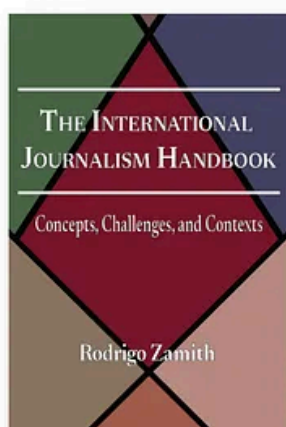
In this article, I hope to show that AI-powered Q&A can be a robust tool for Q&A on complex texts, using more rigorous methods of In-text citing and data structure.

Arguments with citation plays a crucial role in academic writing and in any context where ideas and information are borrowed from other sources. A reliable answer from an AI reading assistant should be based on factual and verifiable information.

For students, academics, analysts, lawyers, etc., checking the source for their continued work is essential. In this article, I aim to showcase a few examples of achieving more reliable responses for your custom documents using in-text citations.

1. Text Processing

For this article, I will show how to chat with the textbook 'The International Journalism Handbook', by Rodrigo Zamith, a college textbook for journalism and media studies. The ebook is under CC BY-NC so it's for any non-commercial use. You can download the file for your desired format.




The International Journalism Handbook

 Rodrigo Zamith

Available Formats (open access)

 [PDF Version](#)

 [Web Version](#)

 [ePub Version](#)

Description

International journalism is crucial to our understanding of the world beyond our own borders. This book is designed to explain key theories and concepts that allow us to understand the general practice of journalism around the world, and to illustrate some of the challenges that arise from practicing journalism in those contexts. It begins by providing a theoretical foundation that helps us understand why international journalism matters and the key forces that shape what it looks like; highlights some of the key challenges to bearing witness to developments, sourcing information, and simply doing 'the job' of journalism; and describes important similarities and differences in how journalism is imagined and performed in different regions of the world.

<https://books.rodrigozamith.com/the-international-journalism-handbook/> | [cc by-nc](#)

To enable the model to capture the exact passages with citation, we need to segment and index the text in a way that we can refer readers for future reading and to be able to conduct fact-checking. For these reasons, I processed the ebook into a CSV file by paragraphs with id.

You can use BeautifulSoup to phrase the text and add an id to each paragraph.

```
def set_data_text(content: str, article_id: int, count: int):

    # set data-name for elements in text

    soup = BeautifulSoup(content, 'lxml')
    xml = soup.recursiveChildGenerator()

    for element in xml:

        # for paragraph tags

        if element.name == 'p':
            element.attrs['index'] = 's_{}_{}'.format(count, article_id)
            count += 1
        ...
        # you can do the same for other tag element according to your text
        ...

    parsed_text = str(soup)

    return [parsed_text, count]
```

2. Install, import, index, and store data:

I will use the [Langchain RAG](#) Q&A implementation with GPT. You can learn more about Langchain Q&A in this [tutorial](#).

```
pip install --upgrade langchain
pip install openai==0.27.8
..

import os

from langchain.chains import RetrievalQA
from langchain.chat_models import ChatOpenAI
from langchain.document_loaders import CSVLoader
from langchain.vectorstores import DocArrayInMemorySearch
from langchain.indexes import VectorstoreIndexCreator
from IPython.display import display, Markdown

....

os.environ["OPENAI_API_KEY"] = <"openai key">
```

```

...

file = 'International_Journalism_Handbook.csv'
loader = CSVLoader(file_path=file)

...

index = VectorstoreIndexCreator(
    vectorstore_cls=DocArrayInMemorySearch
).from_loaders([loader])

from langchain.embeddings import OpenAIEmbeddings
embeddings = OpenAIEmbeddings()

..

docs = loader.load()

..

#add the emmbeded text into a local data storage

db = DocArrayInMemorySearch.from_documents(
    docs,
    embeddings
)

```

2. Index your query and define your retriever.

```

query = "How has the development of technology impacted the way news
organizations collaborate and share information?"

docs = db.similarity_search(query)

retriever = db.as_retriever()

llm = ChatOpenAI(temperature = 0.0)

qdocs = "".join([docs[i].page_content for i in range(len(docs))])

```

3. Set your prompt response for In-text Citation.

To get a cited answer we need to force the model to ground his arguments from the book paragraphs. Note for this part of the prompt.

- *'Summarize articles with citations.'*

- *'For every sentence, you cite the article name.'*
- *'At the end of your summary, Create a sources list of each result you cited, with the article name, author, and link.'*

Interestingly, compelling the GPT model to cite every sentence enhances the user's readability and comprehension and can help eliminate “hallucinations” — a side effect that may cause AI to generate irrelevant or inaccurate information.

```
response = llm.call_as_llm(f"{qdocs} Question: Please answer the question with
For every sentence you write, cite the book name and paragraph number as <id_x
```

```
At the end of your commentary:
```

1. Add key words from the book paragraphs. /
2. Suggest a further question that can be answered by the paragraphs provided.
3. Create a sources list of book names, paragraph Number author name, and a li

```
display(Markdown(response))
```

According to the book The International Journalism Handbook by Rodrigo Zamith, technology has played a significant role **in** shaping today's journalistic work (s_225_100). The development of the printing press allowed **for** the mass distribution of journalism, although it also imposed limitations on the formats that journalistic products could take (s_225_100). The telegraph enabled the development of newswire services **and** facilitated quick transmission of reports **from** remote locations (s_225_100). On the other hand, the proliferation of the telephone allowed reporters to conduct more reporting **from** within the newsroom by directly contacting their sources (s_225_100).

Technological actants have also influenced the way news audiences **and** journalists communicate **with** each other (s_274_100). Platforms like Twitter have made it easier **for** audience members to provide immediate **and** public feedback to journalists, leading to more meaningful **and** direct audience participation (s_274_100). However, this can also result **in** negative forms of participation, such **as** brigading **and** strategic harassment of journalists (s_274_100).

In recent times, journalists are more likely to work **in** teams, collaborate across organizations, **and** involve their audiences **in** various aspects of news production (s_914_100). This shift has moved away **from** the historical practice of journalists working **in** a more solitary fashion (s_914_100).

The accessibility of news content **and** sources has increased significantly, allowing news audiences to have access to a wide **range** of options (s_268_100). This has made it challenging **for** a single journalistic outlet to gain a near-monopoly on audiences (s_268_100). However, a few large organizations

with strong brand recognition can still capture substantial audiences, while smaller journalistic outlets cater to niche audiences and are often considered interchangeable by users (s_268_100).

Keywords: technology, printing press, telegraph, telephone, audience participation, news production, news content accessibility, journalistic outlets.

Further question: How has the evolution of technology impacted the credibility and trustworthiness of journalistic outlets?

Sources:

Book: The International Journalism Handbook

Paragraph numbers: s_225_100, s_274_100, s_914_100, s_268_100

Author: Rodrigo Zamith

Link: <https://books.rodrigozamith.com/the-international-journalism-handbook/>

As you can see, we get a well-structured answer with citations to the original book passages for every argument the AI model generates. The answer is sufficiently comprehensive and can be tracked to its exact sources for review.

Additionally, I've asked the model to create a source list, add some keywords and suggest a further question for readers to try. Altogether, this forces the model to create a more coherent answer.

The same model can provide short answers for fast overview, and can also provide an answer with citation for full review. The cited answer does provide additional knowledge in this case and many others. You can use this methods to track your model accuracy even if you don't want to display the citation. For book, articles and news etc. the need for a cited response is needed in many cases.

The current state of GPT-4 is able to process the data reliably, with the benefits of a clear and coherent answer to any question. With some modification for the prompts, you can achieve different outputs for different use cases. With the new function calling methods you might be able to do more with this kind of structure for passing data as well.

Please note that this article is experimental, and I'm still testing and the ability and quality to cite sources with generative AI. I will be happy to hear responses, so feel free to clap leave a comment or reach out.

AI

Gpt

Langchain

Academic

Prompt




Follow

Written by Yotam Abraham


261 Followers

More from Yotam Abraham

 Build an AI-powered News Reporter in Python Yotam Abraham

Build an AI-powered News Reporter in Python

This article will showcase an experimental information system that utilizes AI to locate, filter and summarize news sources. I will...

Jun 9, 2023  52 Group of students learning on a wooden table Yotam Abraham in UX Collective


Designing NLP features to help students read

A UX research for Natural Language Processing methods, to help students cope with their reading tasks better.

Jul 19, 2020  303



 Three mobile devices presenting content sites with spacial commentary

 Yotam Abraham

Active Reading for Online Articles

*Products and features overview.

Jan 24, 2019  207



See all from Yotam Abraham

Recommended from Medium

Open in app 

Medium



Search



 Ming

Comparing LangChain and LlamaIndex with 4 tasks

LangChain v.s. LlamaIndex — How do they compare? Show me the code!

★ Jan 11 🖱️ 1.2K 💬 9



○ Sandeep Shah

Exploring RAG Implementation with Metadata Filters — llama_index

Exploring Metadata Filters in RAG with Llama-Index: A Practical Guide

★ Mar 16 🖱️ 62 💬 2



Lists

The New Chatbots: ChatGPT, Bard, and Beyond

12 stories · 450 saves

Generative AI Recommended Reading


52 stories · 1320 saves

What is ChatGPT?

9 stories · 425 saves

Natural Language Processing

1667 stories · 1242 saves

 Florian June in Towards AI

The Best Practices of RAG

Typical RAG Process, Best Practices for Each Module, and Comprehensive Evaluation

 Aug 9  759  4



 Nikita Anand

What is RAG(Retrieval-Augmented Generation)?

✦ Apr 27 🖱 108



 Ahmed Besbes in Towards Data Science

3 Advanced Document Retrieval Techniques To Improve RAG Systems

Query expansion, cross-encoder re-ranking, and embedding adaptors

 Jan 16  1.4K  7



 Han HELOIR, Ph.D.  in Towards Data Science

The Art of Chunking: Boosting AI Performance in RAG Architectures

The Key to Effective AI-Driven Retrieval

★ Aug 18  1K  11



See more recommendations