

AI Research Copilot

By Bhavish Pahwa



Problem

1. **Cumbersome literature searches:** Conventional methods of searching for academic information are time-consuming and often yield overwhelming results.
2. **Complex paper reading experience:** Researchers struggle with dense PDF formats, hindering the efficient extraction of knowledge.
3. **Difficulty in staying current:** Researchers face challenges in staying abreast of the latest developments in their field due to information overload and inefficient search methods.
4. **Limited contextual understanding:** Conventional comprehension of research papers may lack the depth and context provided by advanced language models.



Solution

- Interactive Paper Conversations:

Engage in dynamic dialogues with research papers through chat, transforming the reading experience into an interactive conversation.

- Swift Literature Searches:

Utilize Cohere LLMs and Semantic Scholar API to conduct lightning-fast literature searches, streamlining the process of discovering relevant academic information.

- Read Mode in Markdown:

Seamlessly read and comprehend paper PDFs in markdown format, eliminating the challenges associated with complex and incompatible PDF layouts.



Features

1. Enter the respective url of research paper(currently only Arxiv and ACL Anthology urls are supported) and enable read with AI/ chat with the research paper pdf mode powered by co.chat endpoint with documents mode.
2. Conduct literature-review by leveraging the search powered by Cohere co.chat endpoint and using grounding provided by the web connector(from Cohere) and the custom semantic scholar connector (developed and integrated by us).
3. **Read Mode** to convert research paper pdfs to Markdown format for easier reading experience.



Read with AI mode (powered by Cohere)

- 1) We use LLM-Sherpa's LayoutPDFReader to read the respective PDF and convert it into chunks according to the respective sections, tables etc.:- <https://github.com/nlpmatics/llmsherpa>, https://ambikasukla.substack.com/p/efficient-rag-with-document-layout?r=ft8uc&utm_campaign=post&utm_medium=web
- 2) We use Cohere embedding and re-rank models to get the top chunks
- 3) We pass the top chunks as documents to the co.chat endpoint using the documents mode.

Read with AI mode (powered by Cohere)

Read with AI (powered by Cohere)

 This Application currently only works with arxiv and acl anthology web links which belong to the format:-

- 1) Arxiv:- `https://arxiv.org/abs/paper_unique_identifier`
- 2) ACL Anthology:- `https://aclanthology.org/paper_unique_identifier/`


Enter the arxiv or acl anthology url of the paper

`https://arxiv.org/abs/2310.16787`

Chat

 Ask me a question about the research paper

 What is the data provenance initiative

 The Data Provenance Initiative is a large-scale audit of dataset licensing and attribution in AI. The initiative convenes multidisciplinary efforts between legal and machine learning experts to develop tools and standards to trace the lineage of datasets, from their source, creators, series of license conditions, properties, and subsequent use.

The Data Provenance Initiative's goal is to audit popular and widely used datasets with large-scale legal and AI expert-guided annotation. Some of the tools they have developed include the Data Provenance Explorer, a UI allowing practitioners to trace and filter data provenance for popular open-source finetuning data collections, and the Data Provenance Card, which is a

Your question





Literature-Survey

- 1) We study the SemanticScholar API <https://api.semanticscholar.org/api-docs/graph#tag/Paper-Data> and find the respective endpoints and fields we need to query its knowledge base. We request for an API key from the maintainers of semantic scholar.
- 2) We create a REST API using flask and deploy it on pythonanywhere. We create an endpoint(/search) which accepts JSON POST requests which contain the parameter 'query' and return the results as specified by Cohere. Endpoint:- <https://bphigh.pythonanywhere.com/search>
- 3) We register our api with Cohere as a connector
- 4) We build a search system using the web and custom semantic scholar connectors + co.chat endpoint by Cohere to provide better literature-survey capabilities to researchers.

Literature-Survey

Search powered by Cohere and Semantic Scholar

Enter any query you have related to research/academic topics

Limitations of LLMs

- Web
- Semantic Scholar

Search

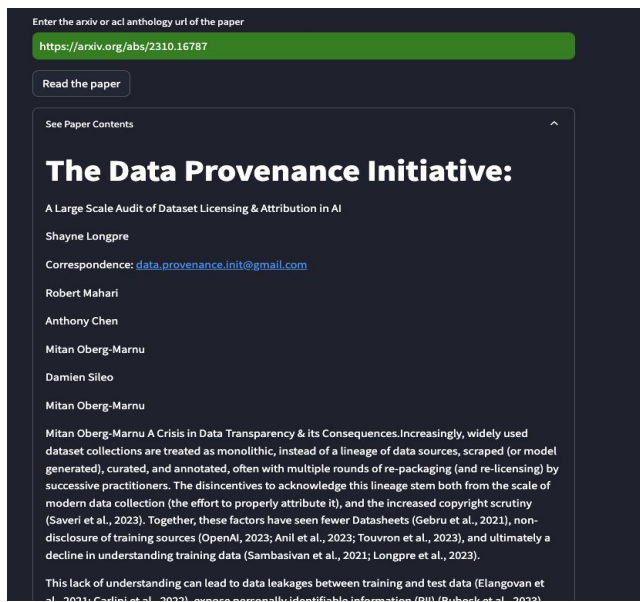
Large Language Models (LLMs) have demonstrated remarkable capabilities in various natural language processing (NLP) tasks. However, they have several limitations, including:

- **Robustness:** LLMs are sensitive to the order of options in multiple-choice questions. This can result in a performance gap of approximately 13% to 75% when answer options are reordered, even in a few-shot setting.
- **Accuracy and appropriateness of generated content:** LLMs have been found to produce inaccurate or inappropriate content. While a methodology of self-correction has been proposed to address this issue, LLMs struggle to self-correct their responses without external feedback, and their performance might even degrade post self-correction.
- **Performance with longer textual content:** LLMs often struggle to identify key information in longer texts, and are more error-prone when summarising longer textual contexts.
- **Limitation of statistical learning:** While LLMs demonstrate that human-like grammatical language can be acquired without a built-in grammar, they still have limitations in explaining the full complexity of human language to cognitive scientists.

It is important to note that LLMs are constantly evolving and improving, and efforts are being made to address these limitations.

Read Mode

1. Enter the respective arxiv or acl research paper url.
2. We use the Meta's recently released Nougat Transformer to convert the research paper to a markdown format:- https://huggingface.co/docs/transformers/model_doc/nougat



The screenshot displays a web interface for reading research papers. At the top, there is a text input field with the URL `https://arxiv.org/abs/2310.16787` entered. Below the input field is a button labeled "Read the paper". Underneath the button is a section titled "See Paper Contents" with a small upward arrow icon. The main content area features the title "The Data Provenance Initiative:" in a large, bold font. Below the title is the subtitle "A Large Scale Audit of Dataset Licensing & Attribution in AI". The authors listed are "Shayne Longpre", "Robert Mahari", "Anthony Chen", "Mitan Oberg-Marnu", and "Damien Sileo". The text of the paper begins with "Mitan Oberg-Marnu A Crisis in Data Transparency & its Consequences. Increasingly, widely used dataset collections are treated as monolithic, instead of a lineage of data sources, scraped (or model generated), curated, and annotated, often with multiple rounds of re-packaging (and re-licensing) by successive practitioners. The disincentives to acknowledge this lineage stem both from the scale of modern data collection (the effort to properly attribute it), and the increased copyright scrutiny (Saveri et al., 2023). Together, these factors have seen fewer Datasheets (Gebru et al., 2021), non-disclosure of training sources (OpenAI, 2023; Anil et al., 2023; Touvron et al., 2023), and ultimately a decline in understanding training data (Sambasivan et al., 2021; Longpre et al., 2023). This lack of understanding can lead to data leakages between training and test data (Elangovan et al., 2021; Casirri et al., 2022), expose personally identifiable information (PII) (Rubsek et al., 2022).

Scope of the product



Total Addressable Market (TAM):

The Total Addressable Market for AI Research Copilot encompasses all individuals and institutions engaged in academic and research activities across various disciplines. This includes universities, research institutions, individual researchers, students, and professionals seeking access to scholarly information. Considering the global nature of academic research, the TAM is expansive and inclusive of diverse fields such as science, technology, engineering, mathematics, humanities, and social sciences.

According to Unesco report:- There were 7.8 million full-time equivalent researchers in 2013, representing growth of 21% since 2007. Researchers accounted for 0.1% of the global population.

Serviceable Addressable Market (SAM):

The Serviceable Addressable Market focuses on the segment of the TAM that the AI Research Copilot can effectively target and serve. This includes researchers and academics who actively engage in literature searches, paper reading, and knowledge extraction. The SAM also extends to institutions and organizations that support and facilitate research activities. As AI Research Copilot offers advanced features such as interactive paper conversations, swift literature searches, and personalized learning experiences, its SAM caters to those who seek to optimize and streamline their research processes.

Scope of the product



Market Scope Overview:

Primary Users:

Individual Researchers: Graduate students, faculty members, and independent researchers looking for efficient tools to enhance their research endeavors. **Academic Institutions:** Universities and research institutions aiming to provide cutting-edge resources to their faculty and students. **Professionals:** Individuals in industries requiring continuous engagement with academic research for innovation and development.

Geographic Scope:

Given the global nature of academic research, the market scope extends internationally, targeting users and institutions around the world.

Industry Focus:

The primary focus is on the academic and research sector, spanning various disciplines, including but not limited to science, technology, engineering, mathematics, humanities, and social sciences.

Competitors



Top competitors are:-

Typeset.io :- <https://typeset.io/>

- Provides paraphraser and AI detection feature but the literature survey tool lacks dedicated knowledge bases

Genie:- <https://www.genei.io/>

Scholarcy:- <https://www.scholarcy.com/>

Future Scope



- 1) In future we would provide the feature of personalized paraphrase where a user can get any text paraphrased according to their own style of writing.
- 2) We are working on a feature to build a social extension to our app so that researchers can connect with each other on the app and can look for people doing work in their areas of interest and collab together.
- 3) Also we would include a remote collab environment where researchers can collab easily using github codespaces and overleaf integrated within a single app to collab in a seamless manner.