AI Assisted Literature Reviews Course

Simon Worthington Renu Kumari Peter Murray-Rust
Gitanjali Yadav Shabnam Barbhuiya Ambreen Hamadani
Moobashara Jawed Anna Rahr Avika Joshi
Deepika Mandakala Anudev Suresh Haarthi Vallabhaneni
Saurav Mishra Anushka Kushwaha Malavika Balachandran
Harshita Mahawar Shaik Zainab

2025-07-22

Table of contents

Using Globally Equitable Open-Source AI LLMs on Open Access Research Literature Repositories

— *For conducting rapid scoping literature reviews*

From: Team #semanticClimate – https://semanticclimate.github.io/p/en/

First release: 2025-07-22

Repo and info | Discussion and support

Course outline: 10.5281/zenodo.16754743 | Software citation information: CITATION.cff

Course units are:

1. Corpus creation: Open Access repository retrieval and analysis
2. Vibe coding: Coding with an AI assistant — Corpus creation
3. PDF summarisation
4. Image classification
5. Image extraction
6. Named entity recognition (NER): Entities and concepts
7. LLM and RAG: For PDF and HTML

You are welcome to use the experimental framework on your own AI assisted literature review. Please get in contact on the discussion board if you have comments or questions.

Additions: A template GitHub repository for running your own AI LLM RAG for a literature review project is provided. It contains software and instructions. The template has been setup for the IPCC AR6 Corpus from the Climate Knowledge Graph but it can be configured to hold any corpus or literature review mini-corpus.

Cite as: 10.5281/zenodo.16779401

URL: https://github.com/semanticClimate/llmrag

The course was run as part of Force11 Scholarly Communications Institute (FSCI) in collaboration with UCLA Library, see course: https://force11.org/fsci/post/fsci-2025-courses-abstracts/#e01 — E01 AI-Assisted Literature Review on Open Access Repositories: Including Image and Object Detection

Thank you to the FSCI organisers and technical support team for the smooth running of the sessions.

Courses dates: Tuesday 2025-7-22, Wednesday 2025-7-23, and Thursday 2025-7-24. Each class will have a Zoom session at the same time on each of the three days, as specified in their course abstracts. The courses are listed in Pacific time (UTC-7)

Abstract

The course covers how to use a self-hosted open-source AI LLM RAG (Retrieval-Augmented Generation) assisted literature review system with supporting user learning material. The system is for: using open access literature repositories; is based on open science (open scholarship) principles; is globally equitable, inclusive, and multilingual, and; is independent of commercial providers.

Participants will be able to self-host their own open-source AI LLM RAG system with no dependency on commercial providers, and to be able to adapt the system to different retrieval and knowledge query use cases. See: semanticClimate llmrag repo.

The Assisted Literature Review (ALR) course covers instruction for a semi-automated literature search with a focus on AI LLM RAG use on a dedicated corpus. In the course the example corpus will be the IPCC's *Sixth Assessment Report*. The framework can be used on any topic or corpus, for example from the Open Access literature from Europe PMC, which is a corpus of 7 million open access articles.

The AI and machine learning open-source software used is the #semanticClimate text and data mining tooling. The course is an introduction to AI Algorithms for data mining including LLM RAG frameworks. A template 'good practice' framework will be provided for participants later use.This course introduces literature search, text mining, image classification as well as object detection. The algorithms and the data used are all open-source and issues of trustability for open science are a priority.

All instruction is carried out using CoLab Jupyter Notebooks so no complicated installations are required.

The learning points covered allow for familiarity with AI tooling for literature search and as a package that can be reused by students and researchers. The learning package already exists as a fully documented workflow, with existing CoLab Notebooks — all deposited in Zenodo with DOIs. The intention is to give participants experience and methodologies to evaluate and integrate AI LLM RAG into their workflows. AI is evolving so fast that focusing on one set of fixed components in a tech stack is not possible, instead the focus is on concept and evaluation.

The focus for the class is a scoping literature review. The results of the AI Assisted Literature Review workflow taught in the class are a literature review report, including: a textual

summary, summaries of papers as a data table, the complete full-text articles downloaded, a reproducible and replicable CoLab Notebook with all the software and code used in the review. The resulting content package can be used in papers, reporting, dashboards, CI pipelines, and for further data analysis.

Contributors

Course chair: Simon Worthington, Climate Knowledge Graph project lead (TIB — Leibniz Information Centre for Science and Technology and University Library) and #semanticClimate member.

Gitanjali Yadav, National Institute of Plant Genome Research (NIPGR) (Co-course chair); Peter Murray-Rust, Cambridge University (Co-course chair); Renu Kumari, National Institute of Plant Genome Research (NIPGR) (Co-instructor).

Additional Contributors: Shabnam Barbhuiya, Jamia Millia Islamia University (Co-instructor); Ambreen Hamadani, Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir (SKUAST-K) (Co-instructor); Moobashara Jawed, Jamia Millia Islamia University (Co-instructor); Anna Rahr, Hannover University of the Applied Science and Arts, and TIB (Co-instructor); Avika Joshi, Delhi Technological University; Deepika Mandakala, Vignan's Institute of Information Technology (A), Visakhapatnam; Anudev Suresh, Jamia Millia Islamia University; Haarthi Vallabhaneni, DVR & Dr. HS MIC College of Technology; Saurav Mishra, National Institute of Technology; Anushka Kushwaha, NIIT University; Malavika Balachandran, University of Toronto; Harshita Mahawar, Amity University, Noida; Shaik Zainab, Anurag University, Hyderabad.

Audience: Researchers, librarians, publishers

Level: (Beginner, but suitable for all levels)

Requirements: Run Google ColLab in a browser. See: https://colab.research.google.com/ A Google account to run CoLab Jupyter Notebooks (if this is not possible users can run Notebooks in their own environments — but please check with course organisers for support). Have a GitLab account (other Git versions can be used — GitLab or Codeberg, etc. Contact course organisers in advance if this is required.) It is also required to obtain some free to use API keys.

Course Learning Objectives

At the end of the course, participants will be able to:

- Conduct a scoping literature review using AI LLM RAG tooling

- Obtain familiarity with using LLM RAG with PDFs and with HTML
- Be able to carry out text and data mining and build a corpus from open access sources such as EPMC
- Use Colab Jupyter Notebooks, execute python commands, and use GitHub.
- Use the provided 'good practice' framework for managing LLM projects

FSCI Learning modules

1. Corpus creation: OA repository retreival and analysis
2. Vibe coding: Coding with an AI assistant — Corpus creation
3. PDF summarisation
4. Image classification
5. Image extraction
6. Named entity recognition (NER): Entities and concepts
7. LLM and RAG: For PDF and HTML

Corpus creation: OA repository retreival and analysis

Renu Kumari

- Notebook – 10.5281/zenodo.16418987
- Output
- Presentation

Vibe coding: Coding with an AI assistant - Corpus creation

Peter Murray Rust

- Slide notes
- Slides
- Documentation
- Code

PDF summarisation

Shabnam Barbhuiya

- Notebook – 10.5281/zenodo.16526790
- Presentation

Image classification

Ambreen Hamadani

- Notebook_Vision Transformers – 10.5281/zenodo.16734915

Image extraction

Avika Joshi

- FigSense – Code: 10.5281/zenodo.16752114

Named entity recognition (NER): Entities and concepts

Moobashara Jawed

- Notebook – 10.5281/zenodo.16559426
- Presentation

LLM and RAG: For PDF and HTML

Shabnam Barbhuiya and Anna Rahr

- Notebook for LLM RAG from PDF/XML – 10.5281/zenodo.16675979
- Repo including Notebook for LLM RAG from HTML – *not shown in the course*

Video recordings for three days (one hour each)

- Day 1: Corpus Creation; Vibe Coding
- Day 2: PDF Summarization; Image classification; FigSense demo — Image extraction
- Day 3: Named Entity Recognition; RAG LLM with PDF-XML

Example literature reviews

#semantiClimate runs an ongoing internship coordinated by NIPGR. Students from the programme carry out literature review on chapters of the IPCC Sixth Assessment Report (AR6). Below are video presentations from the students reporting on their respective literature reviews.

- AR6/WG2/Chapter04 – Water
  - By Ms. Haarthi Vallabhaneni
- AR6/WG2/Chapter08 – Poverty, Livelihood, Sustainable development
  - By Malavika Balachandran
- AR6/WG1/Chapter08 – Water Cycle Changes
  - By Ms. Anushka Kushwaha
- AR6/WG1/Chapter04 – Future Global Climate
  - By Ms. Deepika Mandakala
- AR6/WG2/Chapter-06 – Cities, settlements and key infrastructure
  - By Harshita Mahawar

A climate justice literature review

An example of #semanticClimate tooling being used on the question of Climate Justice, 2024.

As well as supporting step-by-step guide.

10.5281/zenodo.16813353

Worthington, Simon, Renu Kumari, Peter Murray-Rust, Gitanjali Yadav, Shweata N Hegde, and Bhadra Parijat. "Creating the Climate Justice Dictionary — A Step-by-step Guide". semanticClimate, August 12, 2025. https://doi.org/10.5281/zenodo.16813353.

Course materials

Bookmark these Git repository:

- https://github.com/semanticClimate/assited-literature-review and
- https://github.com/semanticClimate/llmrag

Other helpful information

#semanticClimate tools and resources: https://semanticclimate.github.io/p/en/posts/resources/

A software and learning framework for a self-hosted open-source AI assisted literature review, based on — open access, global equity, and open science

2nd July 2025

Links and contact

LLM RAG GitHub repository (version for working with IPCC reports): https://github.com/semanticClimate/llmrag

Mission

The mission is to build a self-hosted open-source AI LLM RAG (Retrieval-Augmented Generation) assisted literature review system with supporting user learning material. The system is for: using open access literature repositories; is based on open science (open scholarship) principles; is globally equitable, inclusive, and multilingual, and; is independent of commercial providers.

Objectives and values

1. To create an AI LLM RAG self-hosted platform and infrastructure framework for literature retrieval and knowledge query.

2. Provide supporting learning resources for AI LLM RAG conceptual models and methods using 'learning-and-understanding-by-doing' — its parts and how it impacts the information landscape.

3. The system and its supporting learning material is designed so that newcomers can enable competences, help understand the principles involved, and how it impacts the information landscape.

4. An approach that implements Global South knowledge participation parity (KPP)*, is global in scope and multilingual.

5. To use only open access literature and FAIR data source for AI LLM use.

6. How to manage and use your own open-source AI LLM platform and infrastructure that is: accessible to all, is designed for digital sovereignty, and is fully open science based from the start, as uses — open source software, and creates FAIR data outputs.

7. To create data sets, such as vector databases, of your own that are fully open source, open licenced, open science in all aspects — including FAIR data.

8. Create a Git based template and methodology for working with AI LLMs. The template allows different technologies to be interchanged or for use in different application use cases — chatbots, for research, literature reviews, text search, derivative publishing, etc.

9. The technical architecture can run on the command line, in Jupyter Notebooks and Google Colab, and as a Streamlit software UI. It acts as a complete pipeline system that goes from raw-text to knowledge and query.

10. To enable the understanding of what are the components needed for an AI LLM system.

11. The system is designed for working with and creating text corpora.

12. Aware of climate change impact issues.

*Analogous to the economics framework of purchasing power parity (PPP) which is used for setting regional pricing of products. For PPP indices see: Eurostat data. PPPs convert different currencies into a common unit which equalizes their purchasing power and eliminates

differences in price levels between economies. (Eurostat) Please get in contact directly for prices.

How to implement the objectives and values

Open

- Open source software licencing
- Open science compliant in all areas and values
- FAIR data principles for data use and data production, e.g., in data production, vector databases
- AI LLM is open source (where possible)
- AI LLM restrictions on republishing content or using content in training? (where possible and if needed)
- Full documentation for reproducibility (where possible)
- Research is open science and open notebook science based from the start
- AI regulation compliant
- Digital sovereignty by design

Open access

- Use open access research literature repositories:
  - Redalyc
  - OpenAlex
  - Europe PubMed Central (Europe PMC)
  - bioRxiv
  - Ukrainian OA repository
  - University of Pretoria

– etc.

Global

- AI and LLM without geo blocking
- Multilingual supporting AI LLM
- Open science knowledge equity values followed — UN Open Science Recommendations

Value proposition

- Support trust in users
- Avoid vendor lockin
- Aim to have higher quality, best of class, literature and data retrieval results
- Enable transparency and reproducibility as much as possible
- Institutional knowledge retention
- Support the knowledge commons and biblio-diversity

LICENSE