

1/21/2026

Attempt 1



In Progress

**NEXT UP: Submit Assignment**

Add Comment

**Unlimited Attempts Allowed**





1/15/2026 to 1/28/2026

## ▼ Details

# Build a Reproducible, Local Research Assistant with RAG + Zotero + LLM

## Assignment Purpose

In this assignment, you will create a personal AI research assistant — a system that helps you semantically search your assigned readings, identify relevant papers, ask research questions, and generate meaningful output (e.g., literature reviews, gaps, hypotheses). This is a tool you will use across this course and your doctoral training. You will combine:

-  Zotero (to store, organize, and verify papers/citations)
-  RAG (Retrieval-Augmented Generation) using Python and ChromaDB
-  LLMs (ChatGPT or Claude) to answer research questions and generate output
-  A metadata mapping tool to track and verify citations

 You are free to use an AI assistant (like ChatGPT or Claude) for help writing or debugging Python code, so long as **you understand what the code does**.

## Why This Is Important:

### 1. Accuracy of Citations

LLMs often generate incorrect, incomplete, or entirely fictional citations. This system ensures your citations are real and grounded in *your* curated library, using Zotero as the final check. This protects your academic integrity and reduces manual errors.

### 2. Efficiency of Literature Review

Instead of reading 50–100 papers manually for every new topic, you will use semantic search to rapidly surface only the most relevant ones — saving dozens of hours per semester.

### 3. Reduced Hallucinations

Because you will feed the LLM only the PDFs selected by your RAG, the LLM's outputs are grounded in actual content — not guesses, not abstracts, and not random web text. This reduces hallucination risk dramatically.

### 4. Reusability and Scaling

Once set up, this system can grow with your research. You can scale it to 1,000+ papers, reuse it for your dissertation, and even use it for group collaborations or co-authored publications. It's a durable skill and tool for your entire career.

### 5. Critical Thinking + Coding Practice

You'll engage both computational thinking (building and customizing a tool) and critical analysis (interpreting outputs, verifying claims, improving workflows). You'll also build confidence in working with Python, AI tools, and automation.



## Tools You Will Use

| Tool             | Purpose                                                                   |
|------------------|---------------------------------------------------------------------------|
| Zotero           | Reference manager to store and verify all papers and citations            |
| RAG Starter Pack | Python code to search your readings by meaning (semantic indexing)        |
| ChatGPT / Claude | Generate lit reviews, gaps, and hypotheses from selected papers           |
| Zotero-to-CSV    | Script to extract metadata from your Zotero library for citation tracking |



## Starter Files (Download Provided)

You will receive one ZIP file:

**Full RAG Assignment.zip** (<https://uta.instructure.com/courses/255203/files/41153842?wrap=1>)   
[https://uta.instructure.com/courses/255203/files/41153842/download?download\\_frd=1](https://uta.instructure.com/courses/255203/files/41153842/download?download_frd=1)

Unzip it and follow the instructions inside. It includes:

- All necessary Python code (RAG Starter Pack)
- Citation tracking script
- Step-by-step instruction file (also below)



## Assignment Instructions — 7 Steps



### Step 1: Create and Organize Your Zotero Library (30–60 min)

#### What to do:

1. Create a free Zotero account at <https://www.zotero.org/> (<https://www.zotero.org/>)
2. Install Zotero desktop and the browser connector
3. Create a Zotero collection called **Seminar Readings**
4. Add the reading list papers (via DOI, PDF, or manual entry)
5. Add Zotero add-in to Word.

#### Why this matters:

Zotero is your **citation source of truth**. It helps you track, verify, and reuse papers across the semester and beyond. Even for working papers without DOIs, Zotero can store author, title, and year.



### Step 2: Download PDFs to a Folder (30–60 min)

#### What to do:

1. Download the same reading list PDFs from your Zotero library or from DOI sources
2. Save them into a local folder:  
Example: `your_project/RAG/zotero_library/`

**Why this matters:**

Zotero's internal folders aren't easily readable by Python. Having a separate, clean folder allows the RAG tool to read and index all files.

---

## ◆ Step 3: Build Your Local RAG System (1–2 hours)

**What to do:**

1. Unzip the starter pack
2. Open the `README.md` and follow the steps:
  - Install Python (if needed)
  - Create a virtual environment
  - Install required packages (`pip install -r requirements.txt`)
  - Place PDFs in the `/zotero_library/` folder
  - Run `build_index.py` to create the semantic index

**Why this matters:**

You are now building a **semantic search engine** that understands the *meaning* of your papers — not just their keywords.

---

## ◆ Step 4: Ask a Research Question (30–45 min)

**What to do:**

1. Run `query.py`
2. When prompted:
  - Enter your research question (e.g. “How does investor sentiment affect asset pricing?”)
  - Choose the number of papers to retrieve (e.g., `top_k = 10`)
  - Optionally set a relevance score threshold (e.g., `min_score = 0.45`)
3. RAG will place the most relevant papers into a folder like `/query_outputs/`.

**Why this matters:**

This step lets you narrow from 70+ papers to the **most relevant 5–15 (or more or less - depends)**. You can adjust how selective RAG is:

- `top_k`: how many papers to return
  - `min_score`: how strict the match must be (try 0.3–0.7)
- 

## ◆ Step 5: Use ChatGPT or Claude to Analyze Papers (30–60 min)

**What to do:**

1. Go to ChatGPT or Claude (Pro access may be required)
2. Upload the PDFs selected by RAG
3. Use the following prompts:

**Prompt A — Literature Review**

Based only on these papers, write a literature review about the relationship between X and Y. Include in-text citations by author and year.

### Prompt B — Research Gaps

Based on these papers, what gaps exist in the literature?

### Prompt C — Hypothesis Development

Propose 1–2 testable hypotheses based only on the uploaded papers.

#### Why this matters:

LLMs can now generate summaries and insights from the right subset of papers — without hallucinations from the open web.

---

## ◆ Step 6: Verify Citations with Zotero (30–60 min)

#### What to do:

1. Check all citations used by the LLM
2. Look them up in Zotero
3. Replace incorrect/missing citations using the correct metadata

#### Why this matters:

LLMs can *still* get citations wrong. Zotero is your tool to verify and correct any citation issues before you submit.

---

## ◆ Step 7: Export Zotero Citation Log (15 min)

#### What to do:







1. Run `zotero_to_csv.py` in the assignment folder
2. This script will export a file like `zotero_paper_map.csv`
3. Review it and include it with your submission

#### Why this matters:

This citation log helps verify the origin of every paper you use and serves as a reproducibility log.

---

## Your Final Submission Should Include:

-  Your original research question(s)
  -  Folder of PDFs selected by RAG
  -  Literature review (as PDF)
  -  Gap identification write-up (PDF)
  -  Hypothesis development write-up (PDF)
  -  Zotero-generated citation log (CSV)
- 

## Notes & Tips

- You may use AI (ChatGPT, Claude) to help you write, troubleshoot Python, or summarize.

- Don't skip verifying citations — this is your scholarly responsibility.
- You can reuse the RAG system all semester for your group project or dissertation work.
- Other tools - like Scispace do a good job explaining difficult papers.

## ✓ Done!

You've now built a **career-long AI assistant** that helps you analyze and extract meaning from hundreds or even thousands of papers.

Absolutely — explicit instructions are essential to prevent the LLM from hallucinating or referencing outside sources.

Below is the **refined and fully explicit prompt template**, designed to force the LLM to work **only with the uploaded PDFs**, with stronger language and structure for accuracy and reproducibility.

## LLM Prompt Template: AI Literature Review Based Only on Uploaded PDFs

### ◆ STEP 1: CONTEXT SETUP (Required First Message)

You are acting as a research assistant for a doctoral student in financial economics.

The user has uploaded a small set of academic papers (PDFs).

These were pre-selected using a local RAG (retrieval-augmented generation) system.

**You must use only the uploaded PDFs** for all answers.

- ✗ Do not include knowledge from outside the uploaded files.
- ✗ Do not infer or hallucinate information not clearly supported in these PDFs.
- ✓ Base all responses on the actual content of the uploaded PDFs.

### ◆ STEP 2: TASK PROMPTS

#### \* Prompt A: Literature Review

Based **only** on the uploaded PDFs:

1. Write a structured literature review on the following research question:

[INSERT RESEARCH QUESTION HERE]

2. Organize the review by key themes, findings, or methodological approaches.
3. Identify consensus, disagreement, or tensions in the literature.
4. Use **in-text citations** in the form (Author, Year) — but **only** for documents among the uploaded PDFs.
5. Do not reference any article unless it is one of the uploaded files.

#### \* Prompt B: Identify Research Gaps

Based **only on the uploaded PDFs**, identify **2–4 research gaps** that remain unresolved or underexplored in the literature. For each gap:

- Briefly explain the gap
- Reference at least one uploaded study that highlights or contributes to the gap
- Do not speculate based on outside sources

### \* Prompt C: Hypothesis Development

Based on your review of the uploaded PDFs and the identified gaps:

- Propose **1–3 testable hypotheses** grounded in the existing literature
- Explain which specific findings from the uploaded PDFs motivate each hypothesis
- Do not include speculative or general hypotheses that are not traceable to the provided documents

### \* Prompt D (Optional): Annotated Bibliography

For each uploaded paper, provide:

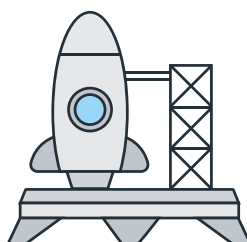
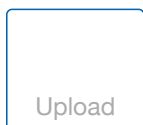
- A brief (3–5 sentence) summary
- Explanation of how it relates to the central research question
- Citation in APA format (but only if the metadata is visible in the PDF — otherwise, indicate that citation details are incomplete)

## 📌 FINAL REMINDER TO ADD:

- ✓ **Reminder:** Work only with the uploaded PDFs.
- ✗ Do not include any paper, idea, author, or finding that is not clearly visible in the uploaded files.
- ⚠️ If citation metadata (author, year, etc.) is not visible in the file, note that clearly.
- 🧠 Your job is to generate grounded, reproducible research outputs based only on this limited library.

Keep in mind, this submission will count for everyone in your Project Group group.

### Choose a submission type



Choose a file to upload

or

 Webcam Photo

 Canvas Files



(<https://uta.instructure.com/courses/255203/modules/items/10714187>)



Assignment

(<https://uta.instructure.com/courses/255203/modules/items/10714187>)