ML2025 Homework 1 - Retrieval Augmented Generation with Agentic System

TA: Ulin Sanga 陳宥林、馮柏翰、劉建豐

Email: ntu-ml-2025-spring-ta@googlegroups.com

Kickoff: 2025/2/28 00:01:00 (UTC+8)

Deadline: 2025/3/21 23:59:59 (UTC+8)

Links

1. Course website: https://speech.ee.ntu.edu.tw/~hylee/ml/2025-spring.php
2. NTU COOL: https://cool.ntu.edu.tw
3. JudgeBoi: https://ml.ee.ntu.edu.tw/home
4. HW1 Google Colab: https://colab.research.google.com/drive/1OGEOSy-Acv-EwuRt3uYOvDM6wKBfSEID?usp=sharing
5. HW1 Kaggle: https://www.kaggle.com/code/u0ulin/ml2025-homework-1

Outline

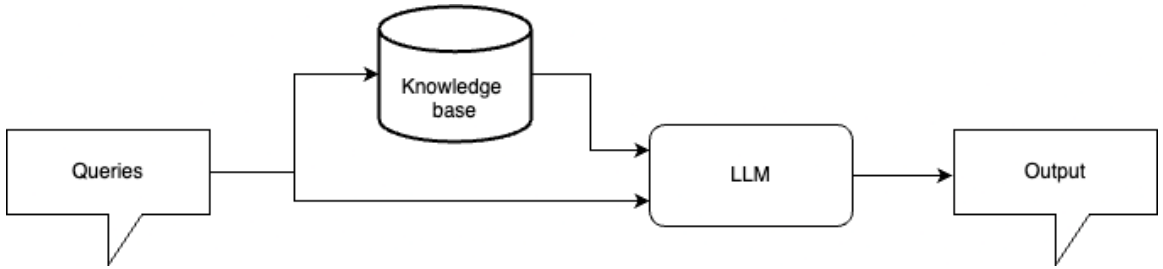
- Task Overview
- TODOs
- Dataset
- Submission and Gradings
- Regulations
- Hints
- References and Appendices

Task Overview

What is RAG?

Retrieval augmeted generation (RAG) is a method that allows LLMs to answer the query with external knowledge.

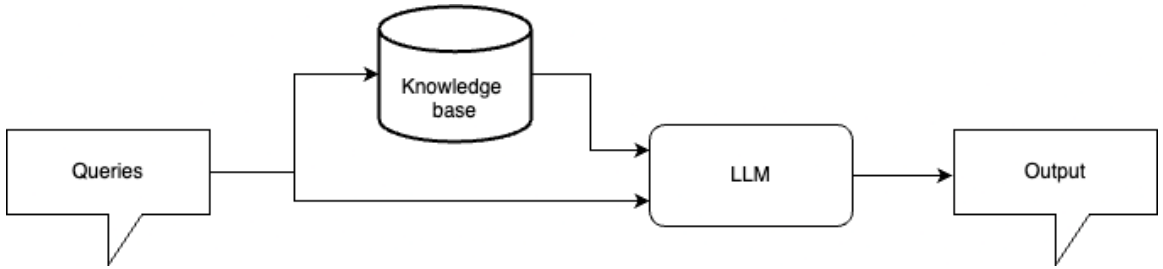
In a naive RAG approach, the query will be fed to a knowledge base to gather relevant information first.



Task Overview

What is RAG?

Then, both the query and the retrieved information will be passed to the LLM simultaneously, allowing the LLM to answer the query with external knowledge.



Task Overview

Why RAG?

- **Knowledge cutoff**
  An LLM must be trained on data that "exist" at the time it was trained.
  For example, an LLM trained in 2024 will not know about anything in 2025.
  RAG can let the LLM have the access to the latest information.
- **Reducing the cost of training**
  When dealing with private data that an LLM unlikely have seen, one can choose to fine-tune the LLM. However, it is costly to fine-tune an LLM. RAG is another approach to allow LLMs to access the private data without the need of training.

Task Overview

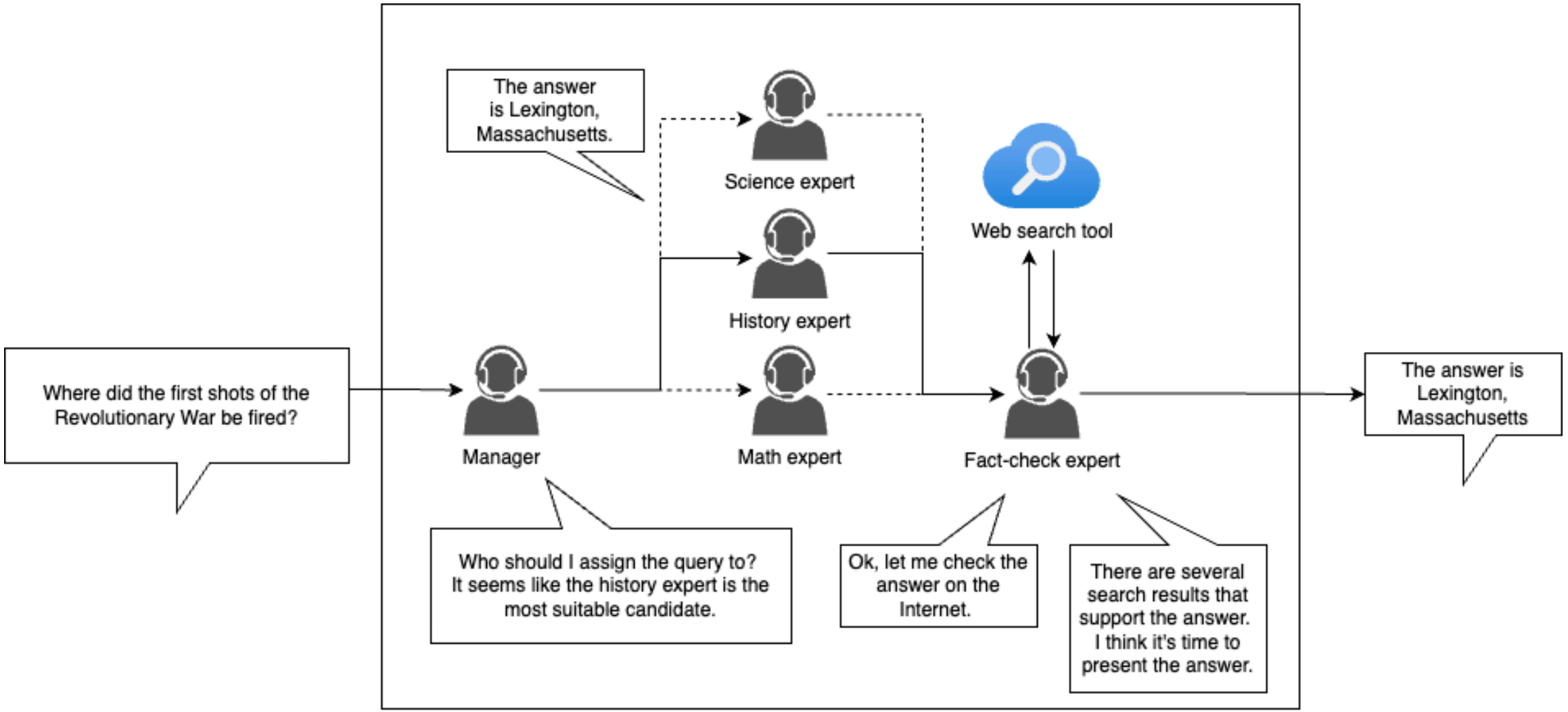
Why RAG?

- **Improving the reliabilities of generated answers**
  By the nature of LLMs, it is a common issue that LLMs will produce hallucinations. With RAG, one can examine the source of the generated texts, hence improve the reliabilities of the answers.

Task Overview

What is Agentic System

An agentic system is a framework where LLMs act as individuals and cooperate to complete complex tasks.


Where did the first shots of the Revolutionary War be fired?
Manager
The answer is Lexington, Massachusetts.
Science expert
History expert
Math expert
Web search tool
Fact-check expert
The answer is Lexington, Massachusetts
Who should I assign the query to? It seems like the history expert is the most suitable candidate.
Ok, let me check the answer on the Internet.
There are several search results that support the answer. I think it's time to present the answer.

Task Overview

RAG with Agentic System

The goal of HW1 is **building an RAG system with agents**.

TODOs

1. Learn how an RAG system operates.
2. Design the prompts for agents to solve some tasks.
3. Submit your code to NTU COOL and submit your results to JudgeBoi.

Dataset

There are 90 hand-crafted questions.

The answers to these questions are a word or a phrase.

The dataset contains two parts: **public** and **private**.

Dataset

Public dataset

For questions in the public dataset, the answers are given.

- Example
  - Q：熊信寬，藝名熊仔，是臺灣饒舌創作歌手。2022年獲得第33屆金曲獎最佳作詞人獎，2023年獲得第34屆金曲獎最佳華語專輯獎。請問熊仔的碩班指導教授為？
  - A：李琳山

Dataset

Private dataset

For questions in the private dataset, the answers are not given.

- Example
  - Q：2005 播出的電視劇《終極一班》中，有一個高中生戰力排行榜，稱為「KO 榜」，該榜榜首為？
  - A：

Submission and Gradings

There are two places to submit!

You have to submit the answers produced by your system to JudgeBoi.
(The link is provided in page 2.)

You will get real-time scores on JudgeBoi.

Submission and Gradings

There are two places to submit!

You have to submit your code to NTU COOL.

- The file name should be [Student ID].zip.
- The zipped file should contains only one file, [Student ID].ipynb/.py, with which we should be able to reproduce one of your selected JudgeBoi submissions.
- All the English alphabets in your student ID should be in lowercase.
- We will only grade your latest submission.

Submission and Gradings

We will use GPT-4o to check if your answers and the standard answers are the same or not.

For example, if the standard answer is 「國立臺灣大學」, you will still get full points if your answer is 「正確答案應該是台大」.

Submission and Gradings

A total of 90 questions and 10 points.

- 4 points: Submit your code to NTU COOL and your results to JudgeBoi.
- 6 points: The correctness of your answers. In both public and private datasets, 1 point for each baseline.

	Public (30 questions)	Private (60 questions)
Simple baseline	4	6
Medium baseline	10	15
Strong baseline	18	22

Do not try to manually modify the results from your system!

Regulations

- You should NOT plagiarize, if you use any other resource, you should cite the source in the comments.
- You should NOT modify your prediction files manually.
- Do NOT share codes or prediction files with any living creatures.
- Do NOT use any approaches to submit your results more than 5 times a day.
- Do NOT use any proprietary models or services that require API keys since this would cause the reproduction process more difficult.

Your final grade x 0.9 and the grade for this homework will be 0 if you violate any of the above rules first time.
Your will get an F for the final grade if you violate any of the above rules multiple times.
Prof. Lee & TAs preserve the rights to change the rules & grades.

Hints

Please note that the following hints are just for your reference. You are not required to follow them. And you are not guaranteed to beat the baselines by simply following them.

- For the simple baseline, you don't need to do anything. Directly feed the questions to the LLM. (Just run the sample code!)
- For the medium baseline, you will need to integrate the search tool into your system. The search tool has already been implemented in the sample code. You need to integrate it into your pipeline.

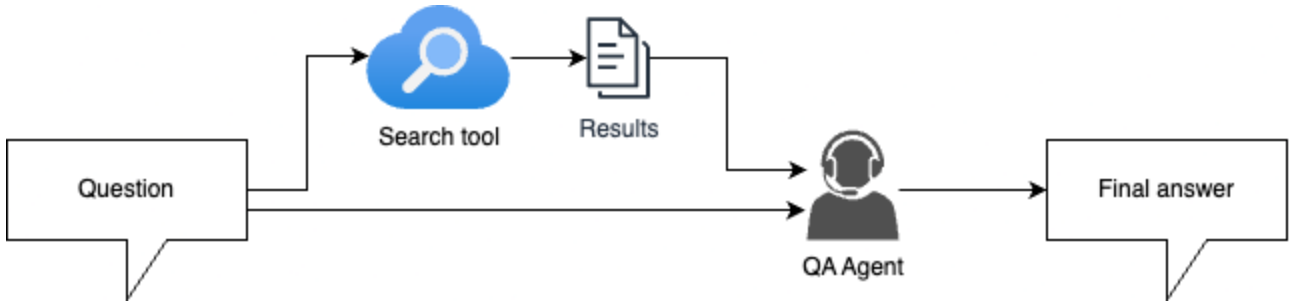
Hints

- For the strong baseline, you need to do some preprocessings on the questions.
- How would you solve this question in the real world? 「熊信寬，藝名熊仔，是臺灣饒舌創作歌手。2022年獲得第33屆金曲獎最佳作詞人獎，2023年獲得第34屆金曲獎最佳華語專輯獎。請問熊仔的碩班指導教授為？」
  - Will you directly paste the question into Google search? Or you will extract the keywords and emphasize on where the question is?
  - To me, I will first search these keywords: 「熊仔 碩班 指導教授」
  - The key question is only 「請問熊仔的碩班指導教授為？」. The rest are irrelevant and can be ignored.
- But how to do the extractions? This is where agents can help! Two agents are enough: Keyword extraction agent, Question extraction agent.

Hints

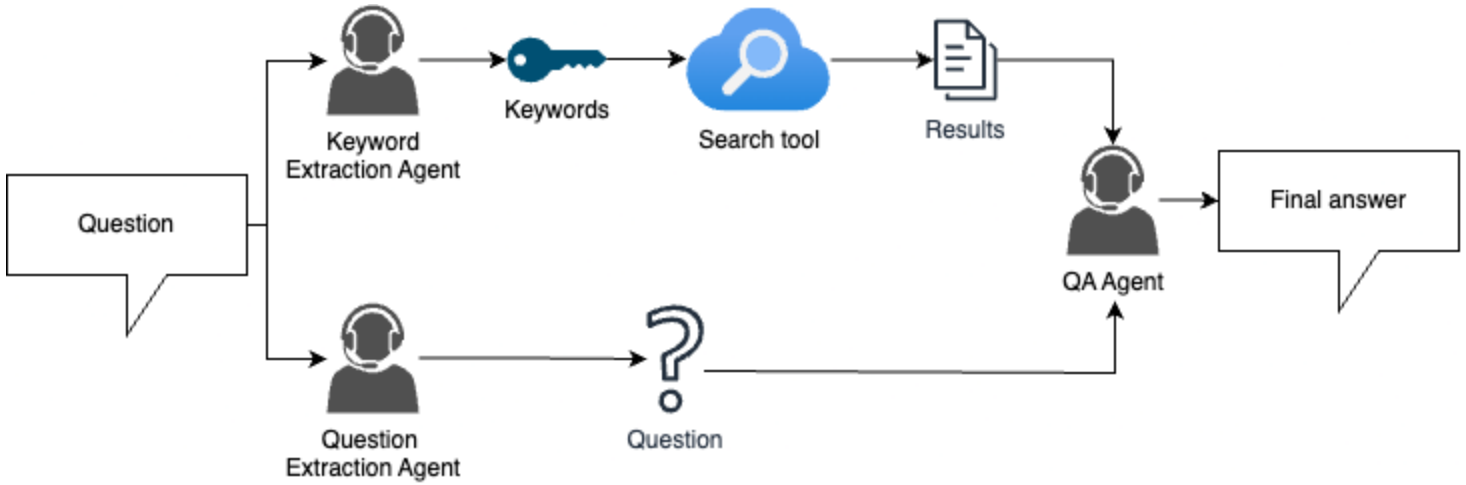
- Simple baseline

- Medium baseline



Hints

- Strong baseline



References and appendices

Prompting techniques:

https://www.youtube.com/watch?v=A3Yx35KrSN0