ADTA 5770: Generative AI with Large Language Models

Thuan L Nguyen, PhD

Semester Project

1. Overview

The final project covers all the topics discussed during the course. The materials posted in any format for the class activities should be considered and used for the project. Additionally, the student can use any other source of information that they can gather.

IMPORTANT NOTES:
--) Students should present their work for each section using text and images.
--) The sources can be from class lectures, assignments, and more, or other sources
--) One picture is worth 1000 words. However, an image with text explaining what it is and what it is for is considered complete and much more convincing.
--) Images can include the screenshots the student has taken while working on the classwork.

IMPORTANT NOTES:
--) When discussing a topic or answering a question, the student must provide adequate explanation and supporting details to their presentation.

IMPORTANT NOTES:
--) If MS Docx is the document format required for submission, the student must submit the contents as MS Docx files, ***not*** *submit PDF documents.*
--) However, before submitting, the student should make a backup copy of the documents by converting them into PDF files that could be used for re-submission if the submitted file was corrupted.

2. Final Project: Assignment Format

The final project is a team assignment, meaning all the group's student members will collaborate while working on it.

However, each student must **write** and **submit** their report **independently**. In other words, a student works on the assignment with the team but **writes** and **submit**s the report as if they had worked independently.

3. Final Project: Overview

Each group is assumed to be an AI system development team in a business organization. With the explosion of popularity and widespread use of generative AI in real-world management and business activities, the corporation’s leaders want the team to develop a generative AI system that the company employees can use to perform content searches, ask questions, and get answers about the contents of the organization’s proprietary documents.

The team will adopt Google Cloud Platform (GCP): Vertex AI services as the primary system Integrated Development Environment (IDE) to design, build, and test the system throughout the project, including but not limited to cloud storage, vector embeddings generation, vector databases management, and advanced vector search technologies. For development, the group will use Python for coding with Google Collaboratory (Colab) as the coding IDE. The group also plans to use popular generative AI techniques, including but not limited to Retrieval Augmented Generation (RAG), Sentence Transformers, and tools provided by generative AI platforms like LangChain and Hugging Face.

The team plans to use Gemini 2.0 Flash and the latest versions of both Python and LangChain in the system development process.

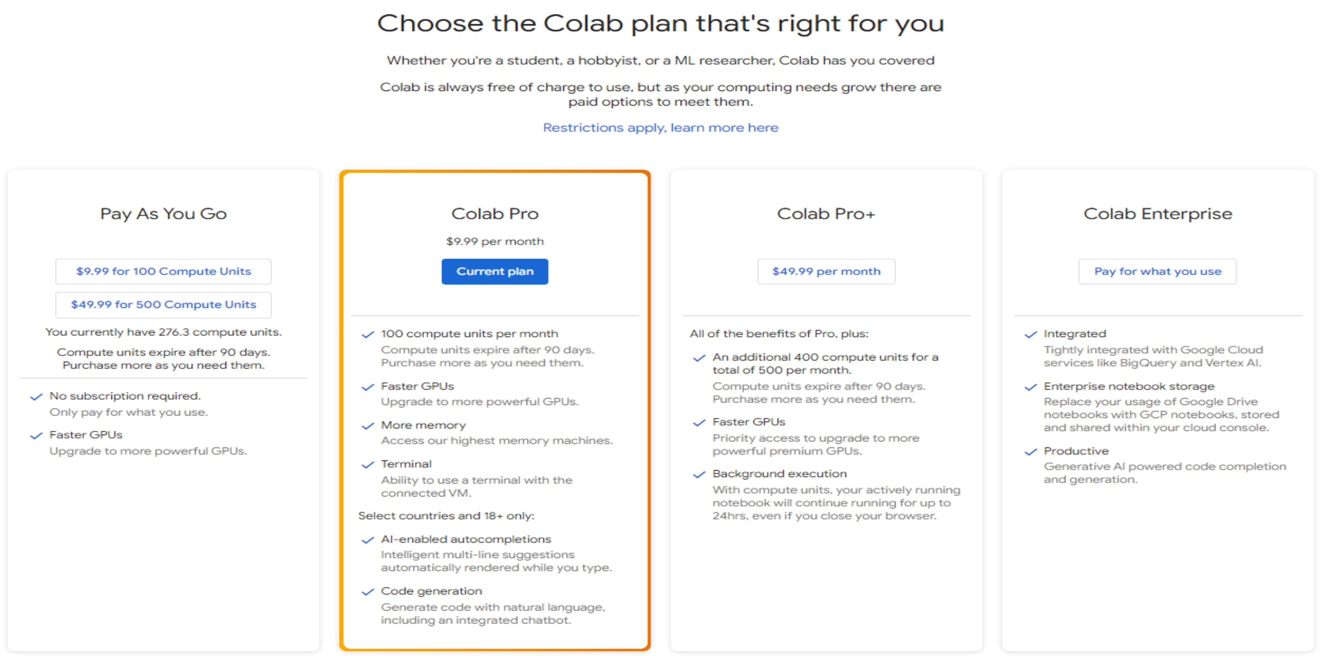
4. PART I: Create a Google Colab Account

TO-DO: Individual students

- Each student must create a free account with Alphabet/Google Collaboratory.

TO-DO: Groups

- Each group must create a paid PRO account with Alphabet/Google Collaboratory.



Choose the Colab plan that's right for you

Whether you're a student, a hobbyist, or a ML researcher, Colab has you covered

Colab is always free of charge to use, but as your computing needs grow there are paid options to meet them.

Restrictions apply, learn more here

Pay As You Go

$9.99 for 100 Compute Units

$49.99 for 500 Compute Units

You currently have 276.3 compute units.

Compute units expire after 90 days. Purchase more as you need them.

- ✓ No subscription required. Only pay for what you use.
- ✓ Faster GPUs Upgrade to more powerful GPUs.

Colab Pro

$9.99 per month

Current plan

- ✓ 100 compute units per month Compute units expire after 90 days. Purchase more as you need them.
- ✓ Faster GPUs Upgrade to more powerful GPUs.
- ✓ More memory Access our highest memory machines.
- ✓ Terminal Ability to use a terminal with the connected VM.

Select countries and 18+ only:

- ✓ AI-enabled autocompletions Intelligent multi-line suggestions automatically rendered while you type.
- ✓ Code generation Generate code with natural language, including an integrated chatbot.

Colab Pro+

$49.99 per month

All of the benefits of Pro, plus:

- ✓ An additional 400 compute units for a total of 500 per month. Compute units expire after 90 days. Purchase more as you need them.
- ✓ Faster GPUs Priority access to upgrade to more powerful premium GPUs.
- ✓ Background execution With compute units, your actively running notebook will continue running for up to 24hrs, even if you close your browser.

Colab Enterprise

Pay for what you use

- ✓ Integrated Tightly integrated with Google Cloud services like BigQuery and Vertex AI.
- ✓ Enterprise notebook storage Replace your usage of Google Drive notebooks with GCP notebooks, stored and shared within your cloud console.
- ✓ Productive Generative AI powered code completion and generation.

5. PART II: Project Proposal (10 Points)

5.1 Overview

To prepare for the semester project, each group needs to submit a project proposal that clearly describes what the group aims to accomplish with the project, including the title, the abstract, and **most importantly**, the **domain expertise field** chosen by the group.

IMPORTANT NOTES:
--) Each group designs, implements, and tests a Q&A Search system using Generative AI technologies.
--) The group aims to deploy the system for generative AI applications applied to a particular real-world business or industry sector, such as engineering, sciences, healthcare, business, and education, to name a few.

- For engineering, the group must specify an engineering area, such as software engineering, chemical engineering, mechanical engineering, etc.
- For business, the group must specify an engineering area, such as accounting, finance, supply chain, etc.
- And so on, …

TO-DO:

- Write a **project proposal** (at least two pages – font-size: 12 – single space – no images)

SUBMISSION REQUIREMENTS PART II:

The student must submit the project proposal by email to the instructor (Thuan.Nguyen@unt.edu)

The email subject: "ADTA 5770: Final Project Proposal - Submission"

Due date & time: 11:00 PM – Wednesday 02/12/2025

6. PART III: Knowledge Base (10 Points)

6.1 Overview

To prepare for the semester project, each group must collect 100 PDFs used as the knowledge base of the Q&A Search system. The document contents must be closely relevant to the domain expertise field chosen by the group.

The instructor has assigned each group a UNT OneDrive folder (via email) to store the knowledge-based documents.

TO-DO:

- Collect 100 PDFs related to the domain expertise field that the group focuses on
- Access the assigned OneDrive folder

6.2 PART III: HOWTO Submit

The group must submit the 100-PDF knowledge base by uploading all the documents to the assigned OneDrive folder.

Then, the group leader must inform the instructor about the submission by email sent to the instructor (Thuan.Nguyen@unt.edu)

The email subject: "ADTA 5770: Final Project Knowledge Base – Group <n> Submission"

Due date & time: 11:00 PM – Sunday 02/16/2025

7. PART IV: Generative AI A&Q-Search System: Planning, Requirements, Data

7.1 Decision on What Type of Business Organization

TO-DO

- Each group is assumed to be an AI system development group in a business organization.
- Brainstorm and discuss with group members to decide on a business organization, such as a company or a corporation, for which the group will develop a generative AI system.

IMPORTANT NOTES:

--) Each group has selected a domain expertise field to focus on while working on the semester project.

--) The group should now place the semester project in the business context of a company or corporation.

SUBMISSION REQUIREMENT'S PART IV #1:
--) Document the group brainstorming and discussion details to decide which type of business organization will develop and deploy the generative AI system.

7.2 Business and Technical Requirements of the System

TO-DO

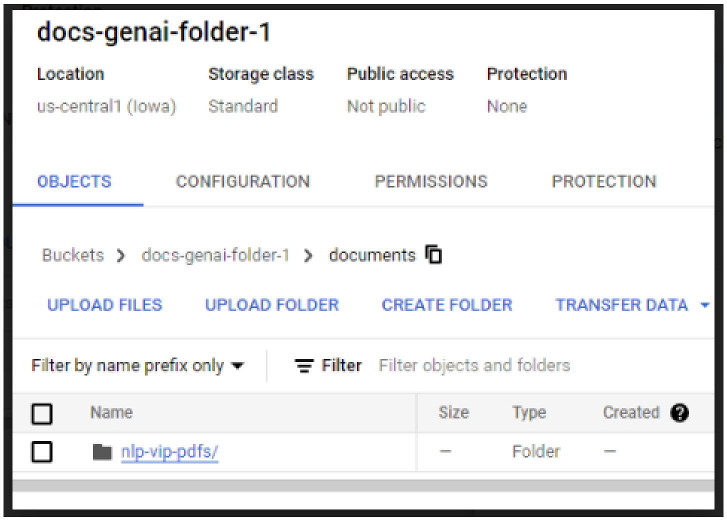
- Based on the business type of the business entity:
  - Document the **business values** of the generative AI system to create
    - Why does the company want to develop a generative AI system?
  - Document the **business requirements** of the generative AI system to create
    - In other words, document the business goals or objectives that the organization tries to achieve with this generative AI system.
- Document the technical requirements of the generative AI system to create.
  - For example:
    - Which AI platform will be used to create the generative AI system?
      - → GCP: Vertex AI
    - Which large language model (LLM)?
    - Which generative AI platform?
    - … and more …

SUBMISSION REQUIREMENTS PART IV #2:
--) Document the details of the business and technical requirements of the generative AI system to be developed.

7.3 Data and Cloud Data Storage Requirements

TO-DO

- Access GCP Storage
- Create one bucket:
  - BUCKET Name = adta5770-docs-folder
- Inside the bucket, create a new subfolder named "documents," in which another subfolder named "pdfs" is created.
- Capture the screenshot of the last subfolder like the following figure:



docs-genai-folder-1

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Not public	None

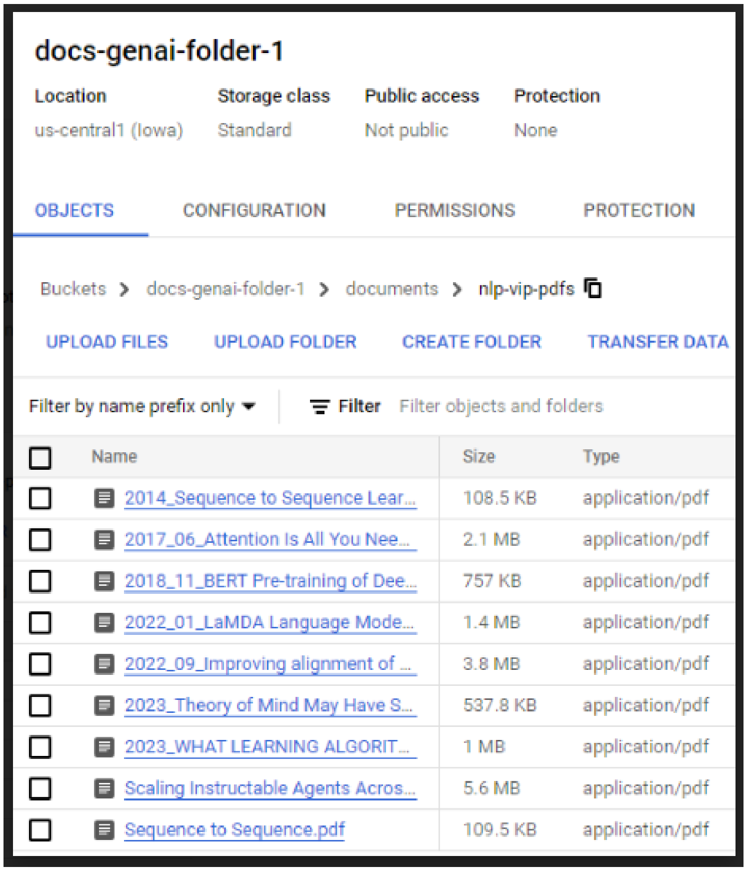
□	Name	Size	Type	Created
□	nlp-vip-pdfs/	–	Folder	–

In the figure, in the bucket docs-genai-folder-1, there exists a subfolder named “documents” in which a sub-sub-folder “nlp-vip-pdfs” is created.

SUBMISSION REQUIREMENTS PART IV #3:
--) Document what has been done in PART III in the HW 4 report.
--) Submit the screenshot like the above for the bucket and its subfolders

7.4 PART IV: Data (PDF Documents) Requirement

- Upload the collected PDFs to the cloud folder: **BUCKET 1/documents/pdfs**
- Access the folder and take a screenshot of its contents like the following one:



docs-genai-folder-1

Location	Storage class	Public access	Protection
us-central1 (Iowa)	Standard	Not public	None

Buckets > docs-genai-folder-1 > documents > nlp-vip-pdfs

Name	Size	Type
2014_Sequence to Sequence Lear...	108.5 KB	application/pdf
2017_06_Attention Is All You Nee...	2.1 MB	application/pdf
2018_11_BERT Pre-training of Dee...	757 KB	application/pdf
2022_01_LaMDA Language Mode...	1.4 MB	application/pdf
2022_09_Improving alignment of ...	3.8 MB	application/pdf
2023_Theory of Mind May Have S...	537.8 KB	application/pdf
2023_WHAT LEARNING ALGORIT...	1 MB	application/pdf
Scaling Instructable Agents Acros...	5.6 MB	application/pdf
Sequence to Sequence.pdf	109.5 KB	application/pdf

SUBMISSION REQUIREMENT: PART IV #4:
--) Document what has been done in PART IV in the HW 4 report.
--) Submit the screenshot as required.

7.5 PART IV: HOWTO Submit

The group must submit all the above-required submissions by uploading all the documents to the assigned OneDrive folder.

Then, the group leader must inform the instructor about the submission by email sent to the instructor (Thuan.Nguyen@unt.edu)

The email subject: "ADTA 5770: Final Project: PART IV – Group <n> Submission"

Due date & time: 11:00 PM – Wednesday 04/02/2025

8. PART V: Generative AI A&Q-Search System: System Analysis

The student is required to submit a project analysis report as the solution to Homework 5.
The report must include the following sections:

1. **Introduction**
   a. Provide an overview of the project.

2. **Problem statement**
   a. Discuss in detail what problem (business, technical, …) the student is trying to solve with this project.

3. **System Requirements Analysis (HW 4)**
   a. Business requirements
      i. Discuss in detail the business requirements of the generative AI system.

   b. Technical requirements
      i. Discuss in detail the technical requirements of the generative AI system.

   c. Data requirements
      i. Discuss in detail the data requirements of the generative AI system.

4. **Feasibility Analysis**
   a. Technical feasibility analysis:
      i. Can we complete the project successfully as required?
      ii. Discuss any technical risks while working on the project.

   b. Business feasibility analysis:
      i. Will the project provide good business value after its completion?
      ii. Discuss any financial risks while working on the project, e.g., running out of funding
   c. Operation feasibility analysis
      i. If we build the system, will it be used by the organization as expected?
      ii. Discuss any risks that may hinder the system's deployment after its completion.

5. **Project Management**
   a. Discuss in detail the timeline of the project
      i. What kinds of significant tasks are to be done?
      ii. What major phases need to be done until the completion?
         1. Due date of each phase or task.

   b. Discuss in detail the human resources needed for the project
      i. How many people, in total, are assigned to work on the project?
      ii. How many people are assigned to work on each major phase of the project?
      iii. For each phase, who does what?

6. **Conclusion**
    a. Provide a short paragraph to express opinions about how the project will be done to conclude the report.
        i. For example: There are many challenges, but they will be successful with great effort.

SUBMISSION REQUIREMENTS: PART V:
--) Submit the system analysis report of the generative AI Q&A-Search system.

8.1 HOWTO Submit

The group must submit all the above-required submissions by uploading all the documents to the assigned OneDrive folder.

Then, the group leader must inform the instructor about the submission by email sent to the instructor (Thuan.Nguyen@unt.edu)

The email subject: "ADTA 5770: Final Project: PART V – Group <n> Submission"

Due date & time: 11:00 PM – Wednesday 04/09/2025

9. PART VI: Generative AI A&Q-Search System: System Design

IMPORTANT NOTES:

*--) To **get credit for PART VI**, the student **must take notes of the lecture (in class) on Wednesday 03/24/2024** and **use the notes to design** the Q& –Search system developed with the project.*

- *To complete this section, the student **cannot** use any content or materials obtained from the Internet or any external source.*

9.1 High Level Design

TO-DO

- Design (high-level) the semester project Q&A-Search system for the semester project

SUBMISSION REQUIREMENT: PART VI #1:

- Submit the high-level design of the Q&A-Search System for the semester project

9.2 Detailed Design

TO-DO

- Design (detailed) the Q&A-Search system for the semester project

SUBMISSION REQUIREMENT: PART VI #2:

- Submit the detailed design of the Q&A-Search System for the semester project

9.3 HOWTO Submit

The group must submit all the above required submissions by uploading all the documents to the assigned OneDrive folder.

Then, the group leader must inform the instructor about the submission by email sent to the instructor (Thuan.Nguyen@unt.edu)

The email subject: "ADTA 5770: Final Project: PART VI – Group <n> Submission"

Due date & time: 11:00 PM – Wednesday 04/09/2025

10. PART VII: Generative AI Q&A-Search System: System Set-Up (CODING – PHASE 1)

TO-DO

- Use Google COLAB as the coding Integrated Development Environment (IDE) to develop code to build the system
- Use Google Cloud Platform (GCP) as the Cloud Integrated Development Environment System (CIDES) to set up (install) all software platforms, tool libraries, and utility software programs needed to build, test, and run the Q&A Search system
- Execute the code, debug, and fix errors if necessary

SUBMISSION REQUIREMENT: PART VII:

- Complete the code in COLAB to set up the software application foundation that can be used to build, test, and run the Q&A Search system
- The code must be executed successfully.

HOWTO SUBMIT

The group must complete the code, run the code successfully, and have it ready for in-class review

Due date & time: 6:00 PM – Monday 04/14/2025

11. PART VIII: Generative AI Q&A-Search System: System Development (CODING – PHASE 2, 3, and 4)

TO-DO

- Use Google COLAB as the coding Integrated Development Environment (IDE) to develop code to build the system

- Use Google Cloud Platform (GCP) as the Cloud Integrated Development Environment System (CIDES) to develop the code to build the Q&A Search system
    - PHASE 2: Import all necessary software application platforms, tool libraries, and utility software applications needed to develop the code for the Q&A Search system
    - PHASE 3: Develop the code (COLAB) to initialize the GCP: AI Platform for the Q&A Search system
    - PHASE 4: Create an empty vector search index and deploy it with a public endpoint

- Execute the code, debug, and fix errors if necessary

SUBMISSION REQUIREMENT: PART VIII:

- Complete the code in COLAB to set up the software application foundation that can be used to build, test, and run the Q&A Search system
- The code must be executed successfully.

HOWTO SUBMIT

The group must complete the code (PHASE 2, 3, and 4), run the code successfully, and have it ready for in-class review

Due date & time: 6:00 PM – Monday 04/14/2025

12. PART IX: Generative AI Q&A-Search System: System Development (CODING: PHASE 1 – 10)

TO-DO

- Use Google COLAB as the coding Integrated Development Environment (IDE) to develop code to build the system

- Use Google Cloud Platform (GCP) as the Cloud Integrated Development Environment System (CIDES) to develop the code to build the Q&A Search system for the phases: PHASE 5 – 10.

- Make a copy of the Jupyter Notebook with the code of PHASE 1, 2, 3, and 4 that has run successfully, say Notebook_2.

- Open **Notebook_2**
- Update the code of PHASE 4 (in Notebook_2) by adding the NEW provided code pf PHASE 4 sent via email

  NOTES:
  --) In Notebook_2: In PHASE 4, the code provides the vector search index ID and its public endpoint ID. Both have been created (by the code of PHASE 4 in the first notebook) and exist.

- Add the code of PHASE 5, 6, 7, 8, 8, and 10
  - The code that each group has taken the photos in the previous class meetings when I discussed the code of each PHASE.

  NOTES:
  *--) In Notebook_2: For PHASE 5 – 9, **each group must have the code available (for grading)** in the notebook.*
  --) The code might or might not run successfully 100%. It's OK.

SUBMISSION REQUIREMENT: PART IX:

- Complete the code in COLAB as required.

HOWTO SUBMIT

The group must submit the Jupyter Notebook with the code of the entire system (PHASE 1 – 10) as above required to the OneDrive group folder.

After the group has submitted the notebook, the group leader must inform the instructor about the submission by emailing the instructor (Thuan.Nguyen@unt.edu).

Due date & time: 11:00 PM – Sunday 04/20/2025

13. PART XI: Generative AI Q&A-Search System: System Testing & Running

TO-DO

- Use Google COLAB as the coding Integrated Development Environment (IDE) to test the complete (10 PHASES) code of the system

- Run **one or two prompts** on the knowledge base to get responses

SUBMISSION REQUIREMENT: PART XI:

- Run the Q&A Search system with prompts to ask questions and get responses.

HOWTO SUBMIT

The group must complete the code (PHASE 1 - 10), run the code successfully, and have it ready for the final in-class design and code review

Due date & time: 8:00 AM – Monday 04/28/2025

(>>>>>>>>>>>>>> Continue the next page >>>>>>>>>>>>)

14. PART XII: Generative AI Q&A-Search System: Prompts and Responses

TO-DO

- Run 10 prompts and get responses on the selected domain field using the Q&A Search system.

SUBMISSION REQUIREMENT: PART XII:

- The submission is a part of the Group Submission Portfolio
    - See Group Submission Portfolio for all the details

Due date & time: 8:00 AM – Monday 04/28/2025