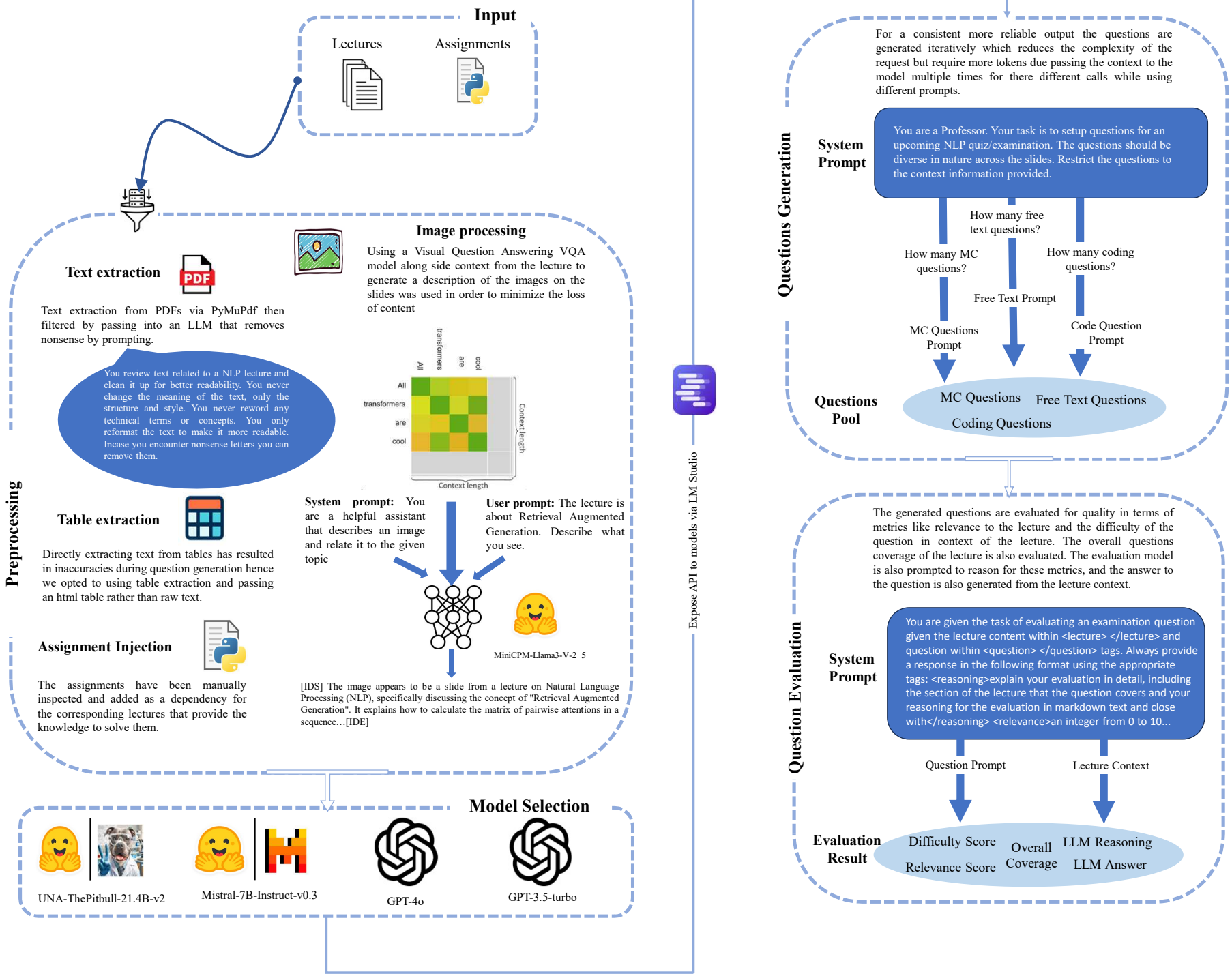
Input
Lectures
Assignments
Preprocessing
Text extraction
PDF
Text extraction from PDFs via PyMuPdf then filtered by passing into an LLM that removes nonsense by prompting.
You review text related to a NLP lecture and clean it up for better readability. You never change the meaning of the text, only the structure and style. You never reword any technical terms or concepts. You only reformat the text to make it more readable. Incase you encounter nonsense letters you can remove them.
Table extraction
Directly extracting text from tables has resulted in inaccuracies during question generation hence we opted to using table extraction and passing an html table rather than raw text.
Assignment Injection
The assignments have been manually inspected and added as a dependency for the corresponding lectures that provide the knowledge to solve them.
Image processing
Using a Visual Question Answering VQA model along side context from the lecture to generate a description of the images on the slides was used in order to minimize the loss of content
All
transformers
are
cool
Context length
System prompt: You are a helpful assistant that describes an image and relate it to the given topic
User prompt: The lecture is about Retrieval Augmented Generation. Describe what you see.
MiniCPM-Llama3-V-2_5
[IDS] The image appears to be a slide from a lecture on Natural Language Processing (NLP), specifically discussing the concept of "Retrieval Augmented Generation". It explains how to calculate the matrix of pairwise attentions in a sequence…[IDE]
Model Selection
UNA-ThePitbull-21.4B-v2
Mistral-7B-Instruct-v0.3
GPT-4o
GPT-3.5-turbo
Expose API to models via LM Studio
Questions Generation
For a consistent more reliable output the questions are generated iteratively which reduces the complexity of the request but require more tokens due passing the context to the model multiple times for there different calls while using different prompts.
System Prompt
You are a Professor. Your task is to setup questions for an upcoming NLP quiz/examination. The questions should be diverse in nature across the slides. Restrict the questions to the context information provided.
How many MC questions?
How many free text questions?
How many coding questions?
MC Questions Prompt
Free Text Prompt
Code Question Prompt
Questions Pool
MC Questions
Free Text Questions
Coding Questions
Question Evaluation
The generated questions are evaluated for quality in terms of metrics like relevance to the lecture and the difficulty of the question in context of the lecture. The overall questions coverage of the lecture is also evaluated. The evaluation model is also prompted to reason for these metrics, and the answer to the question is also generated from the lecture context.
System Prompt
You are given the task of evaluating an examination question given the lecture content within <lecture> </lecture> and question within <question> </question> tags. Always provide a response in the following format using the appropriate tags: <reasoning>explain your evaluation in detail, including the section of the lecture that the question covers and your reasoning for the evaluation in markdown text and close with</reasoning> <relevance>an integer from 0 to 10...
Question Prompt
Lecture Context
Evaluation Result
Difficulty Score
Relevance Score
Overall Coverage
LLM Reasoning
LLM Answer

Q1 What is the importance of adding special tokens like [CLS], [PAD], [SEP], and [MASK] during the preprocessing of text data for transformer models.
Q2 What is a Retrieval Augmented Generator (RAG) system, and how does it help in reducing the context size for generation tasks?
Q3 Explain the concept of "Zero Shot Prompting" and provide an example of a zero-shot prompt scenario.

Q4 What is the purpose of using Term Frequency – Inverse Document Frequency (TFIDF) in NLP?

- To count occurrences of words in a document
- To normalize over documents and make embeddings that facilitate document distinction
- To calculate the number of documents a term appears in
- To determine the total number of terms in a corpus

Q1 Explain how one-hot encoding works and provide an example using a simple sentence.
Q2 How do transformers solve the problem of parallelization in sequence-to-sequence models and why is this significant for NLP tasks?
Q3 What is the primary purpose of training language models with human feedback?

Q4 Discuss the techniques used for handling unknown words in statistical language models.

- To make the models align with their users' preferences
- To reduce the cost of manual data labeling
- To increase the complexity of large prompt datasets
- To automate the process of fine-tuning language models

Q1 How does prefix tuning differ from parameter-efficient fine-tuning methods like LoRa, in terms of their approach to updating parameters while maintaining model efficiency?
Q2 What is the intuition behind the smoothing techniques in statistical language modeling, and how do they help with the sparsity issue of n-gram models?
Q3 Why would a term with an inverse document frequency of 0 not be useful for distinguishing between documents in this context?

Q4 What is the role of the Byte Pair Encoding (BPE) token learner algorithm in text preprocessing for NLP?

- It separates punctuation from words.
- It normalizes case folding, making everything lowercase or uppercase as needed.
- It learns subword tokens that can represent frequent word subparts and are often morphemes
- It performs full stemming of the words in a corpus.

asdasdasdsadsad