

Introduction to LLM

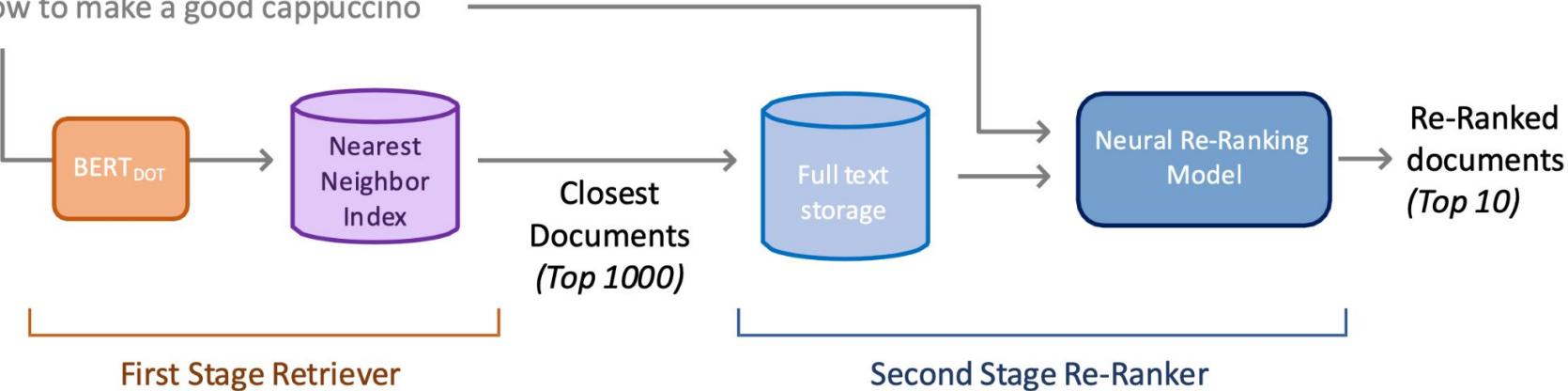
Practice Session 13

Rerankers, RAG, and Agents

High-level Retriever + Reranker Pipeline

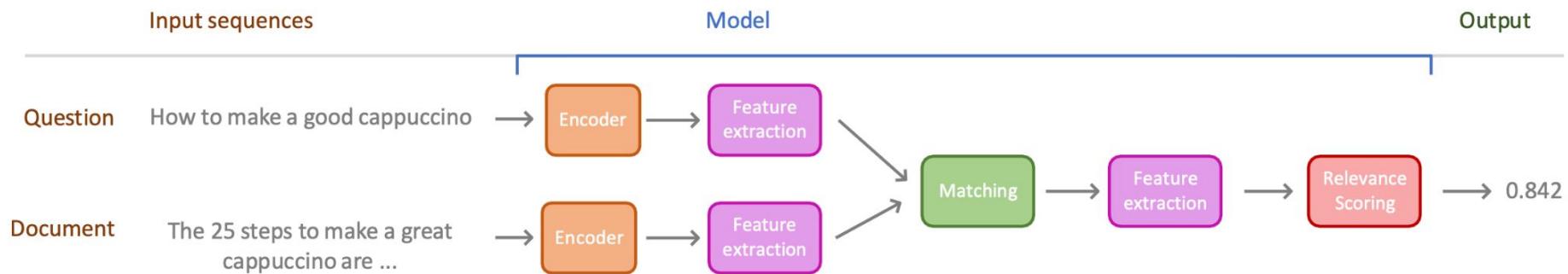
Query

How to make a good cappuccino



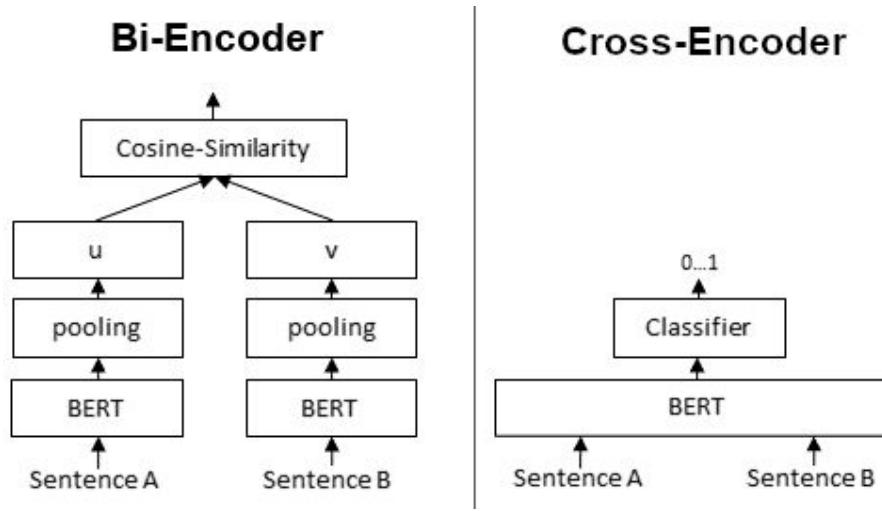
- First stage retrievers can be:-
 - Dense retrievers, e.g., BERT-DOT
 - Sparse retrievers, e.g., BM25
 - Hybrid retrievers (combination of both)

Reranker Main Idea (Interaction Modelling)



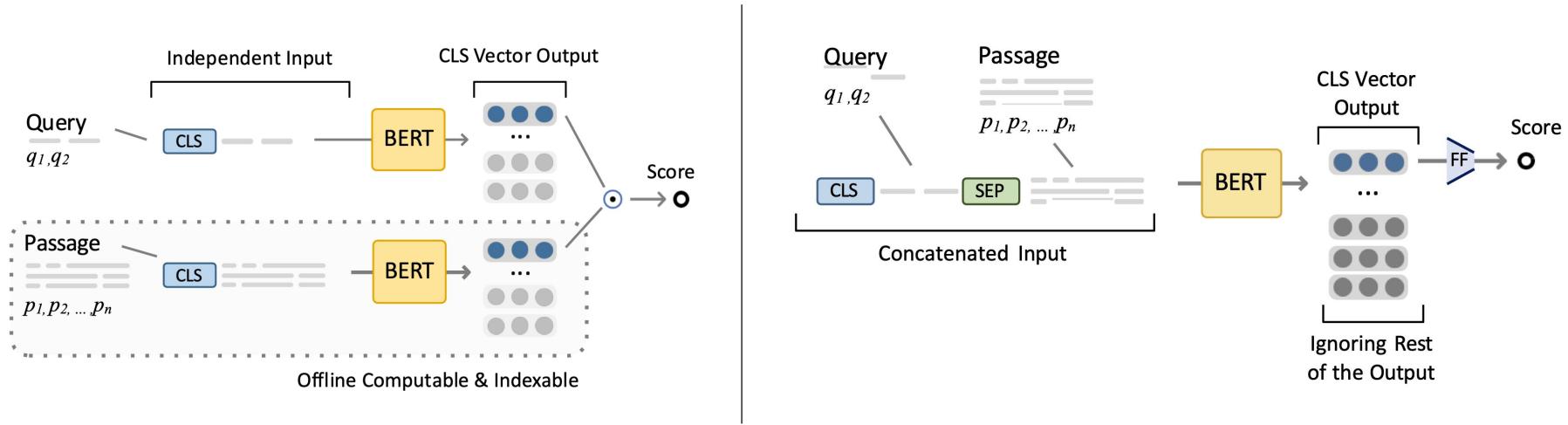
- Rerankers model the interaction between the query and the document
 - Dense retrieval treat the query encoding and document encoding separately (no interaction)
 - Reranker learns to generate one encoding for both query and document together (interaction)
 - This one encoding for both is now used to generate one scalar relevance score

Bi-encoder (Dense Retriever) vs Cross-encoder (Reranker)



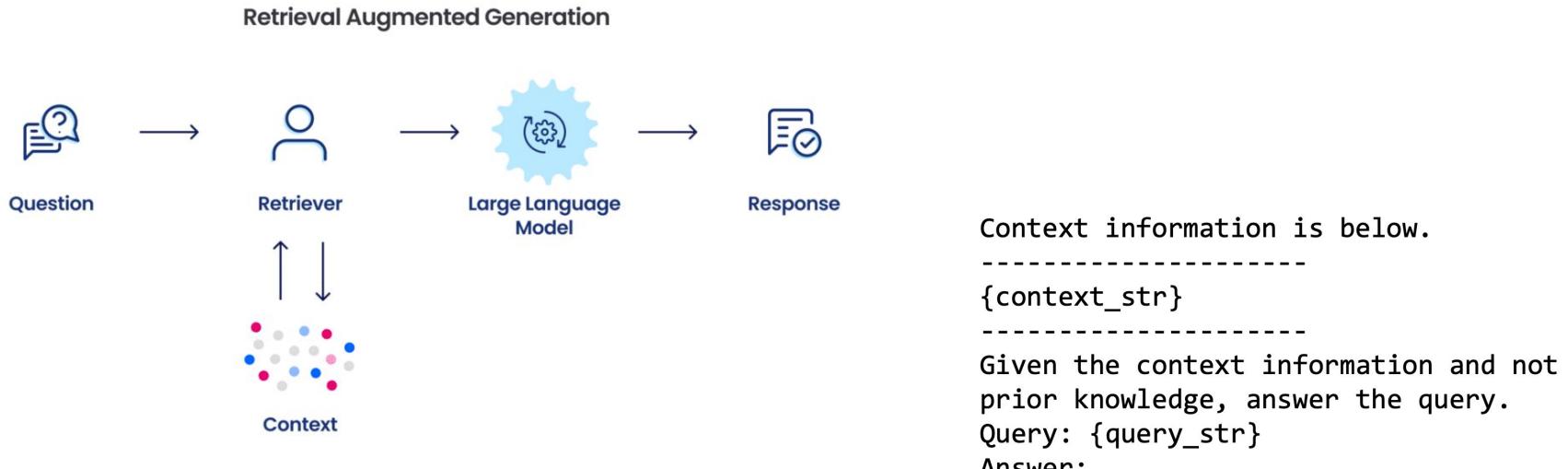
- Source: [Article Link](#)
- Cross-encoders takes both the query (sentence A) and the document (sentence B) simultaneously
 - It learns a classifier (simple NN) to generate one output (score) between 0 and 1

Bi-encoder (BERT-DOT) vs Cross-encoder (BERT-CAT)



- Rerankers (Cross-encoder) is slower as it must be done online (at query time)
 - It's better than bi-encoders in detecting relevance between query and document
 - It acts as a verifier that checks only the top k (e.g., 100) documents from bi-encoders and re-rank them

Retrieval-Augmented Generation (RAG)



- Source: [Article Link](#)
- Step 1: retrieve top N documents using some Retriever pipeline.
- Step 2: write the query with the top N documents (context) for the LLM.

RAG (Interleaving Decomposition) for Distraction Mitigation

Is it true that Colonel Walter Phelps served the United States Army for more than 30 years?

Decomposition step #1



Q1: Who is Colonel Walter Phelps?

Retrieval step #1



E1: Walter Phelps: (Oct 29, 1832–February 20, 1878) was an officer in the Union Army throughout the American Civil War, serving as commanding officer of the Eastern Iron Brigade.

Decomposition ans. #1



E1: Walter Phelps: (Oct 29, 1832–February 20, 1878) was an officer...

Q1: Who is Colonel Walter Phelps?

A1: Colonel Walter Phelps was an officer in the Union Army throughout the American Civil War.

Decomposition step #2



A1: Colonel Walter Phelps was an officer in the ...

Q2: How long did Colonel Walter Phelps serve the United States Army?

Retrieval step #1



E2: Walter Phelps: [['Walter Phelps Jr.'], ['Allegiance', 'United States of America Union'], ['Service/branch', 'United States Army Union Army'], ['Years of service', '1861-1865'], ['Rank', 'Colonel Bvt. Brigadier General']]

Decomposition ans. #2



...

A1: Colonel Walter ...

Q2: How long did Colonel Walter Phelps serve the United States Army?

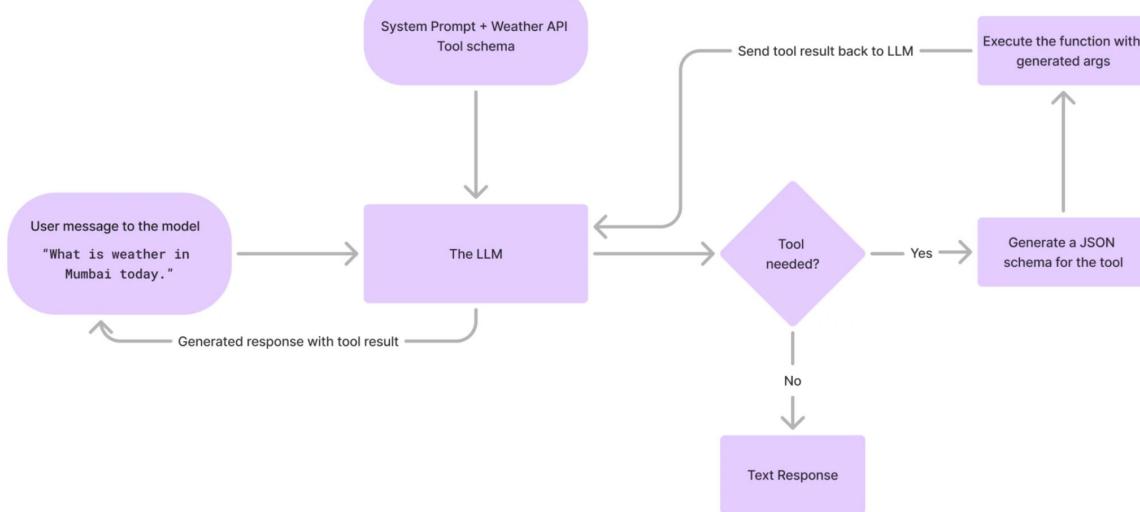
A2: Colonel Walter Phelps served the United States Army for 4 years.

Final answer.



The final answer is:
No.

Tool Use with LLMs



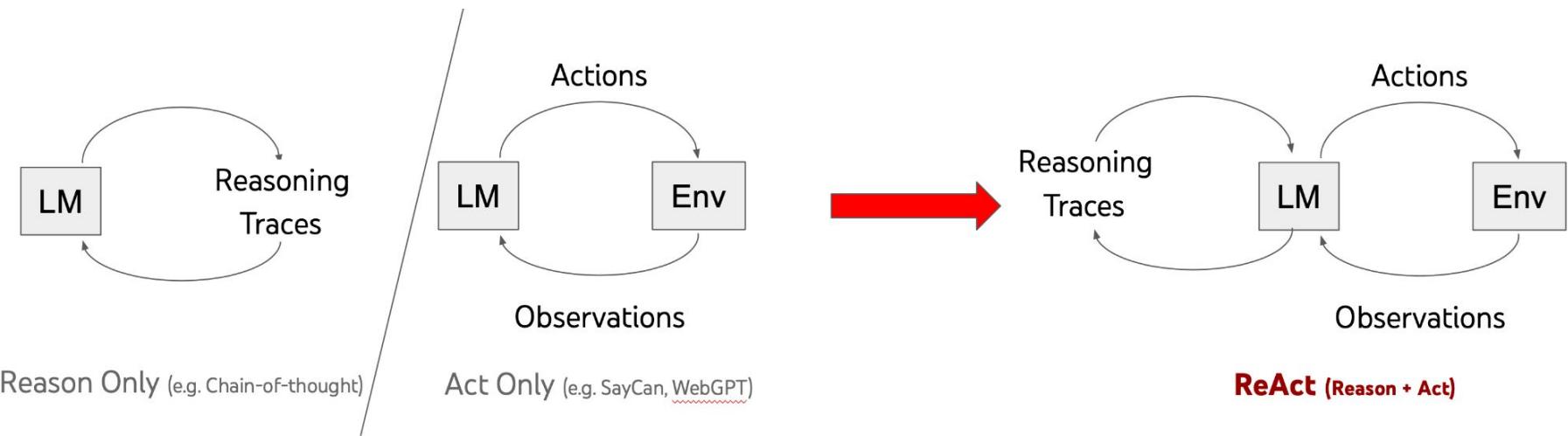
- Source 1: [Article Link](#), Source 2: [Article Link](#)
- LLMs are fine-tuned on numerous tool usage examples (mostly in JSON)

def add(a: int, b:int) -> int:
 """Adds two integers together"""\n return a + b

{
 "type": "function",
 "function": {
 "name": "add",
 "description": "Adds two integers together",
 "strict": true
 },
 "parameters": {
 "type": "object",
 "required": [
 "a",
 "b"
]
 },
 "properties": {
 "a": {
 "type": "integer",
 "description": "The first integer to add"
 },
 "b": {
 "type": "integer",
 "description": "The second integer to add"
 }
 }
}
"additionalProperties": false

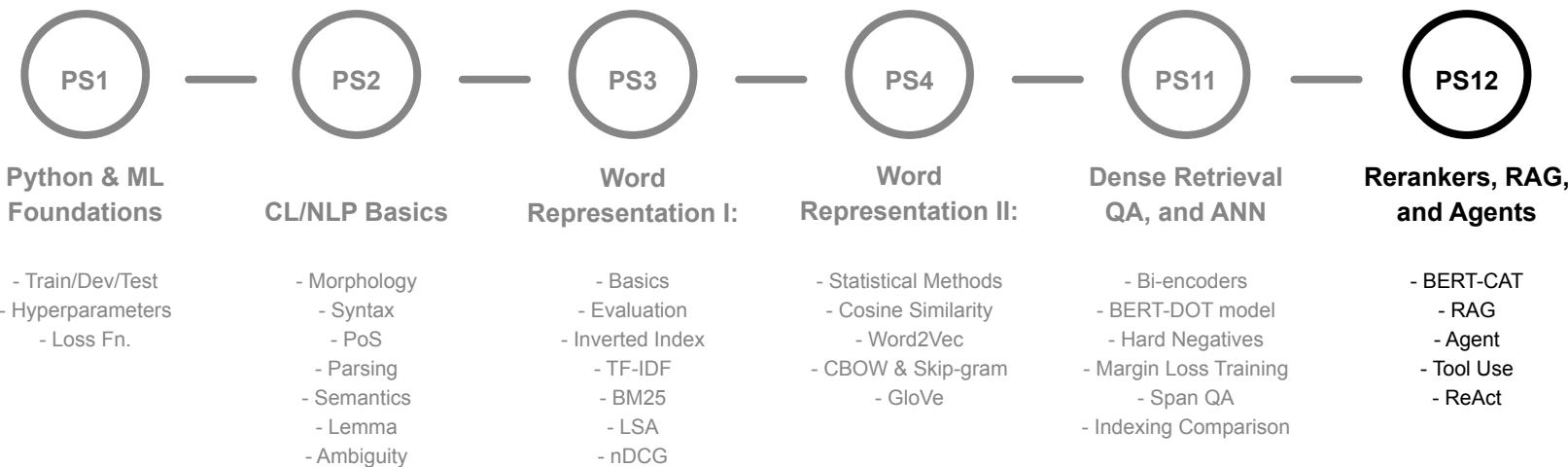
Tool Schema (can be simple string)

ReAct: Synergizing Reasoning and Acting in Language Models

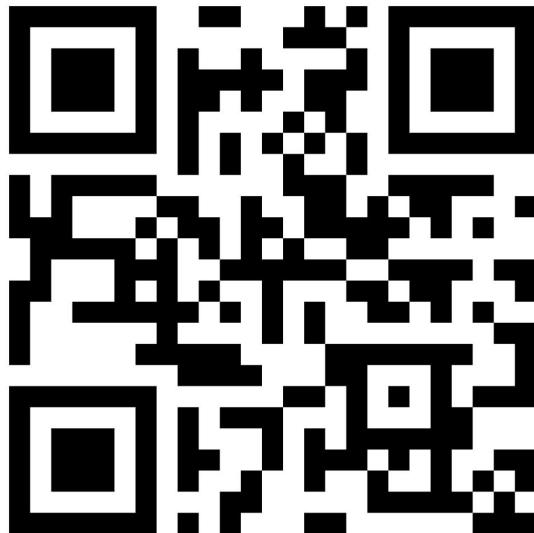


- Source 1: [Article Link](#)
- LLMs are strong at reasoning, but weak at exact arithmetic calculations, or answering based on up-to-date information (training data has a cut-off date) -> Solution: use both thinking + actions (tools)

Timeline



PS13: Colab Notebook (Available on Moodle)



The screenshot shows a Google Colab interface with the following details:

- Title:** Copy of Practice_Session_01_Student_Exercises.ipynb
- Status:** Changes will not be saved
- Toolbar:** File, Edit, View, Insert, Runtime, Tools, Help
- Search Bar:** Commands, Code, Text, Run all
- Copy to Drive button:** A dashed arrow points to this button in the toolbar.
- Table of Contents:**
 - PS 01: Introduction (Python and ML Foundations)
 - Learning Objectives
- Description:** By the end of this practice session, you will be able to:
- Objectives List:**
 1. Know Python basics: variables, data types, operators, and control structures
 2. Manipulate strings effectively: indexing, slicing, and built-in string methods
 3. Work with data structures: lists, dictionaries, and their operations
 4. Handle file I/O: reading/writing text files and JSON data
 5. Use essential libraries: NumPy for numerical computing, Pandas for data manipulation
 6. Apply object-oriented programming: classes, methods, and type hints
 7. Implement basic ML workflows: data splitting, model training with PyTorch

- https://colab.research.google.com/drive/1CQHvt18Y_tujmPTNGeLxLrgzmAxijyvM
- Before running any cell, please choose T4 GPU by clicking on "Runtime" in the main menu then on "Change runtime type", and disconnect from it when you finish using it.