Financial Analysis Using Large Language Models

Jia Finance Practicum Project

Master of Science in Financial Engineering, University of Illinois at Urbana-Champaign. Authors: Ben Granados, Yi Yan, Dhruv Oza, Yunze Liao

Who is JIA Finance?

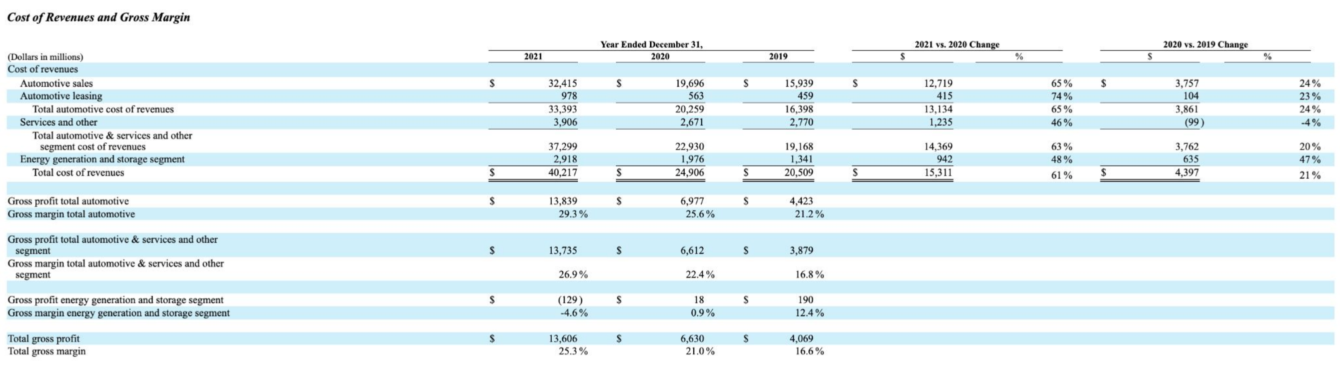
Corporate Advisor: Ishan Prasad

- Jia Finance is a fintech platform that provides mortgages to foreigners purchasing residential real estate in the US.

- The firm utilizes an asset-based approach to underwriting and uses the latest data analysis technology as well as AI-based identity verification to quickly make underwriting decisions to fund mortgages for international buyers.

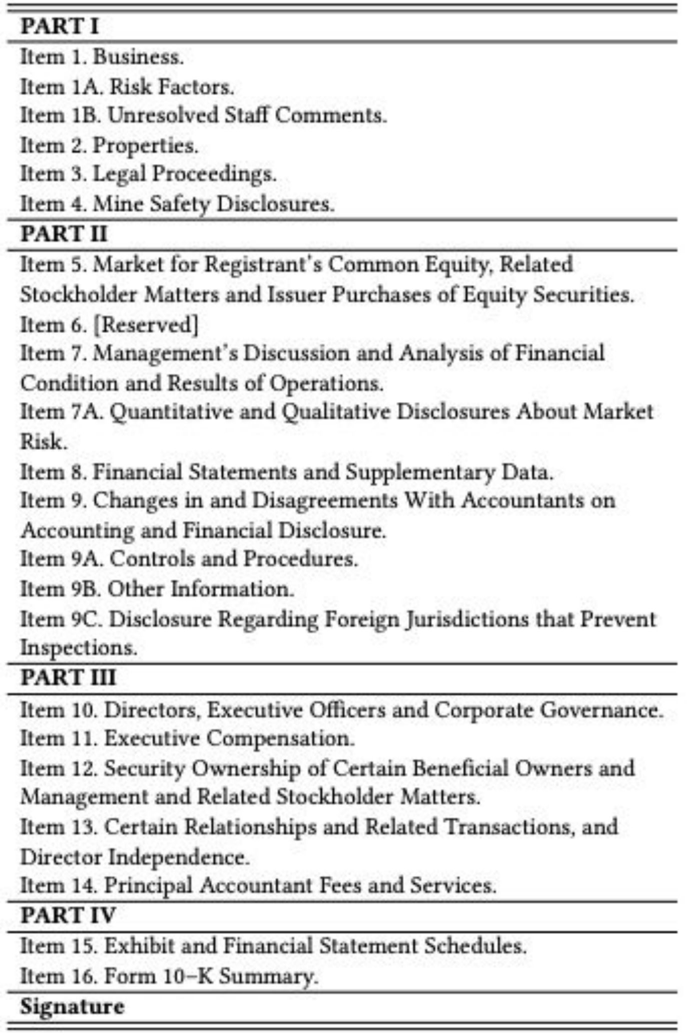
Project Scope

- Using LLM, we can scrutinize financial documents like 10-K forms for tracking financial performance of public companies.
- Though the process of analyzing 10-K requires considerable man hours to sift through the documents. Utilizing natural language processing, investors can swiftly discern the report's tone, extract useful metrics, monitor red flags and keywords, and create benchmark using peer performance.



Cost of Revenues and Gross Margin

	Year Ended December 31,			2021 vs. 2020 Change		2020 vs. 2019 Change	
(Dollars in millions)	2021	2020	2019	$	%	$	%
Cost of revenues							
Automotive sales	$ 32,415	$ 19,696	$ 15,939	$ 12,719	65 %	$ 3,757	24 %
Automotive leasing	978	563	459	415	74 %	104	23 %
Total automotive cost of revenues	33,393	20,259	16,398	13,134	65 %	3,861	24 %
Services and other	3,906	2,671	2,770	1,235	46 %	(99)	-4 %
Total automotive & services and other segment cost of revenues	37,299	22,930	19,168	14,369	63 %	3,762	20 %
Energy generation and storage segment	2,918	1,976	1,341	942	48 %	635	47 %
Total cost of revenues	$ 40,217	$ 24,906	$ 20,509	$ 15,311	61 %	$ 4,397	21 %
Gross profit total automotive	$ 13,839	$ 6,977	$ 4,423				
Gross margin total automotive	29.3 %	25.6 %	21.2 %				
Gross profit total automotive & services and other segment	$ 13,735	$ 6,612	$ 3,879				
Gross margin total automotive & services and other segment	26.9 %	22.4 %	16.8 %				
Gross profit energy generation and storage segment	$ (129)	$ 18	$ 190				
Gross margin energy generation and storage segment	-4.6 %	0.9 %	12.4 %				
Total gross profit	$ 13,606	$ 6,630	$ 4,069				
Total gross margin	25.3 %	21.0 %	16.6 %				



PART I

Item 1. Business.
Item 1A. Risk Factors.
Item 1B. Unresolved Staff Comments.
Item 2. Properties.
Item 3. Legal Proceedings.
Item 4. Mine Safety Disclosures.

PART II

Item 5. Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities.
Item 6. [Reserved]
Item 7. Management's Discussion and Analysis of Financial Condition and Results of Operations.
Item 7A. Quantitative and Qualitative Disclosures About Market Risk.
Item 8. Financial Statements and Supplementary Data.
Item 9. Changes in and Disagreements With Accountants on Accounting and Financial Disclosure.
Item 9A. Controls and Procedures.
Item 9B. Other Information.
Item 9C. Disclosure Regarding Foreign Jurisdictions that Prevent Inspections.

PART III

Item 10. Directors, Executive Officers and Corporate Governance.
Item 11. Executive Compensation.
Item 12. Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters.
Item 13. Certain Relationships and Related Transactions, and Director Independence.
Item 14. Principal Accountant Fees and Services.

PART IV

Item 15. Exhibit and Financial Statement Schedules.
Item 16. Form 10-K Summary.

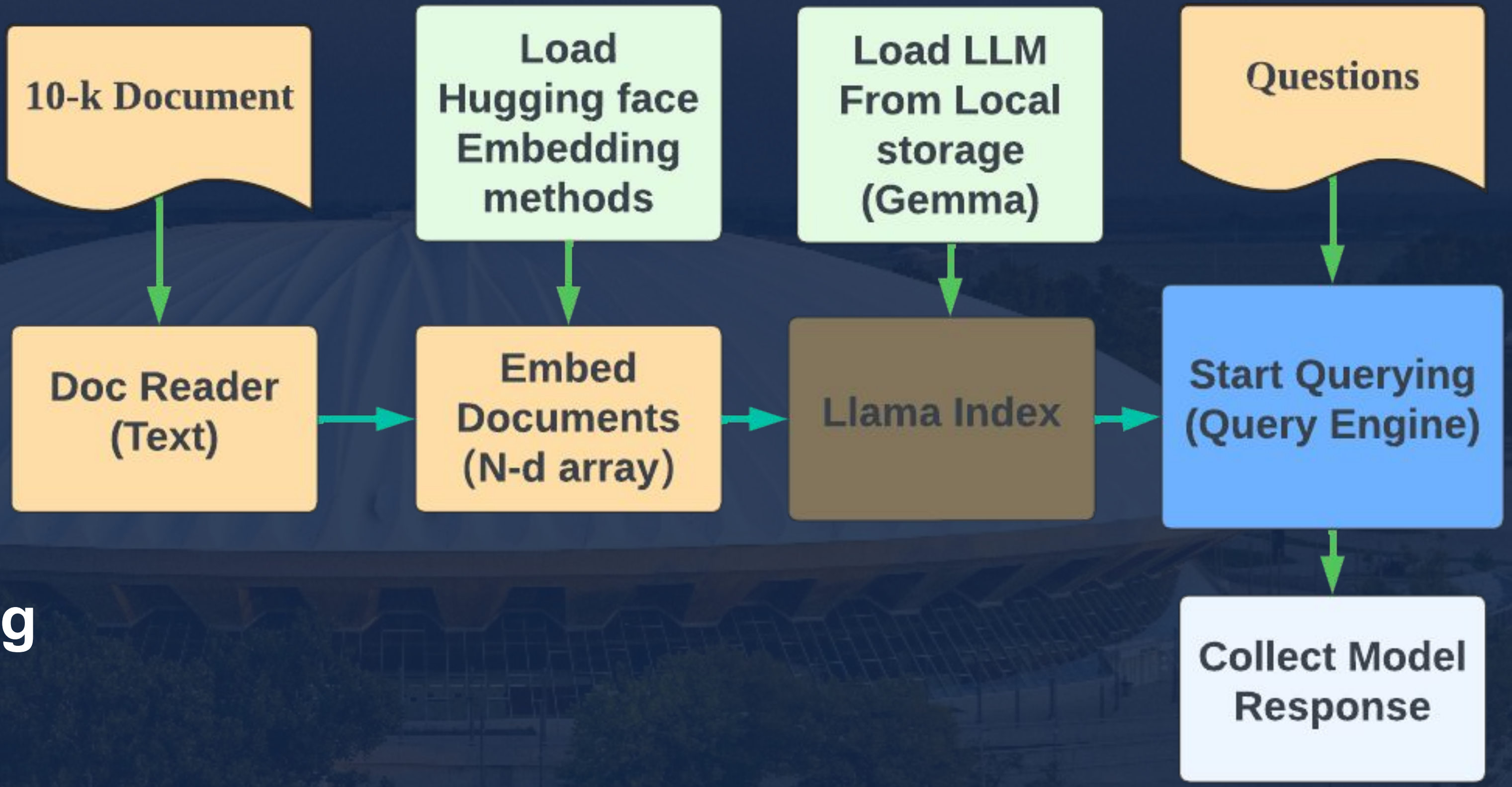
Signature

DEMO

Development Process

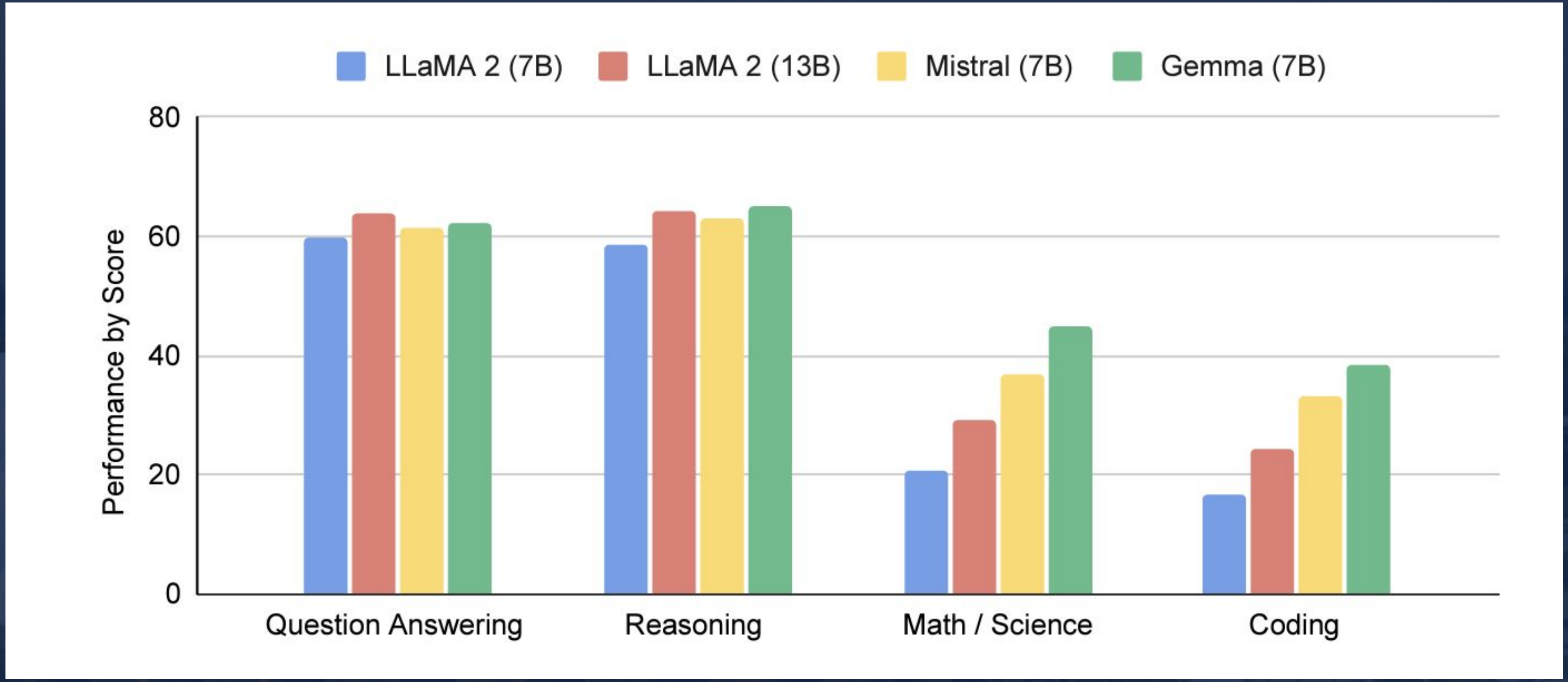
Development Process:

- Preprocessing the Data
- Embedding the Data
- Building the Retriever
- Evaluating and Fine-tuning


10-k Document
Load Hugging face Embedding methods
Load LLM From Local storage (Gemma)
Questions
Doc Reader (Text)
Embed Documents (N-d array)
Llama Index
Start Querying (Query Engine)
Collect Model Response

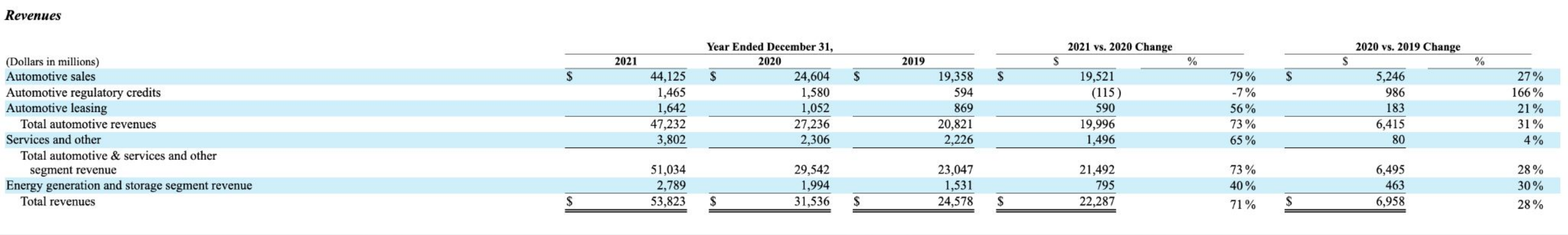
LLM Selection

- Overall choice – **Gemma-7B**
  - based on its performance, efficiency, accessibility, and trainability.
- Model deployment:
  - Security and stability
  - Storage & Memory
    - model size - 30GB
  - Computation
    - running model
    - fine-tuning model
- NCSA's **Delta** cluster.
  - high performance computing cluster offered by U of I
  - 256 GB RAM, 4*A100 GPUs



MMLU benchmark

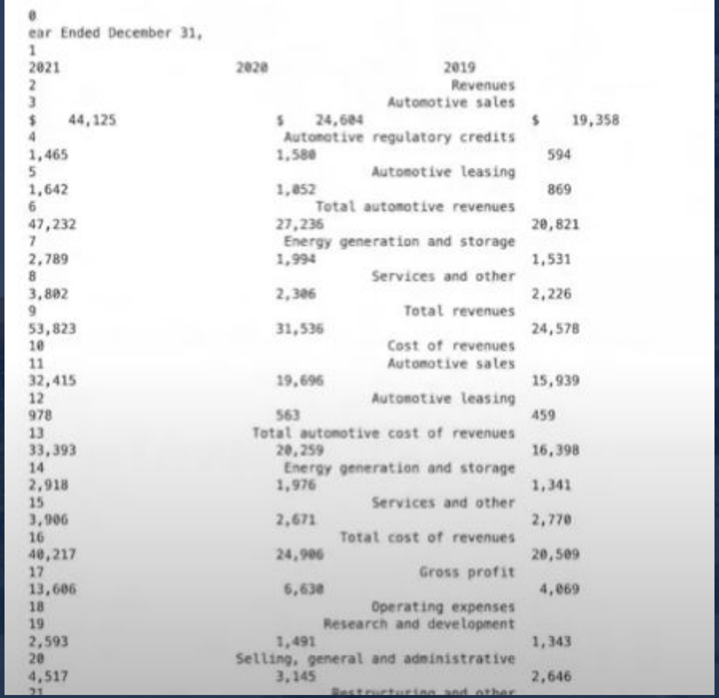
Preprocess Data with Unstructured.io



Revenues

(Dollars in millions)	Year Ended December 31,			2021 vs. 2020 Change		2020 vs. 2019 Change	
	2021	2020	2019	$	%	$	%
Automotive sales	$ 44,125	$ 24,604	$ 19,358	$ 19,521	79 %	$ 5,246	27 %
Automotive regulatory credits	1,465	1,580	594	(115)	-7 %	986	166 %
Automotive leasing	1,642	1,052	869	590	56 %	183	21 %
Total automotive revenues	47,232	27,236	20,821	19,996	73 %	6,415	31 %
Services and other	3,802	2,306	2,226	1,496	65 %	80	4 %
Total automotive & services and other segment revenue	51,034	29,542	23,047	21,492	73 %	6,495	28 %
Energy generation and storage segment revenue	2,789	1,994	1,531	795	40 %	463	30 %
Total revenues	$ 53,823	$ 31,536	$ 24,578	$ 22,287	71 %	$ 6,958	28 %

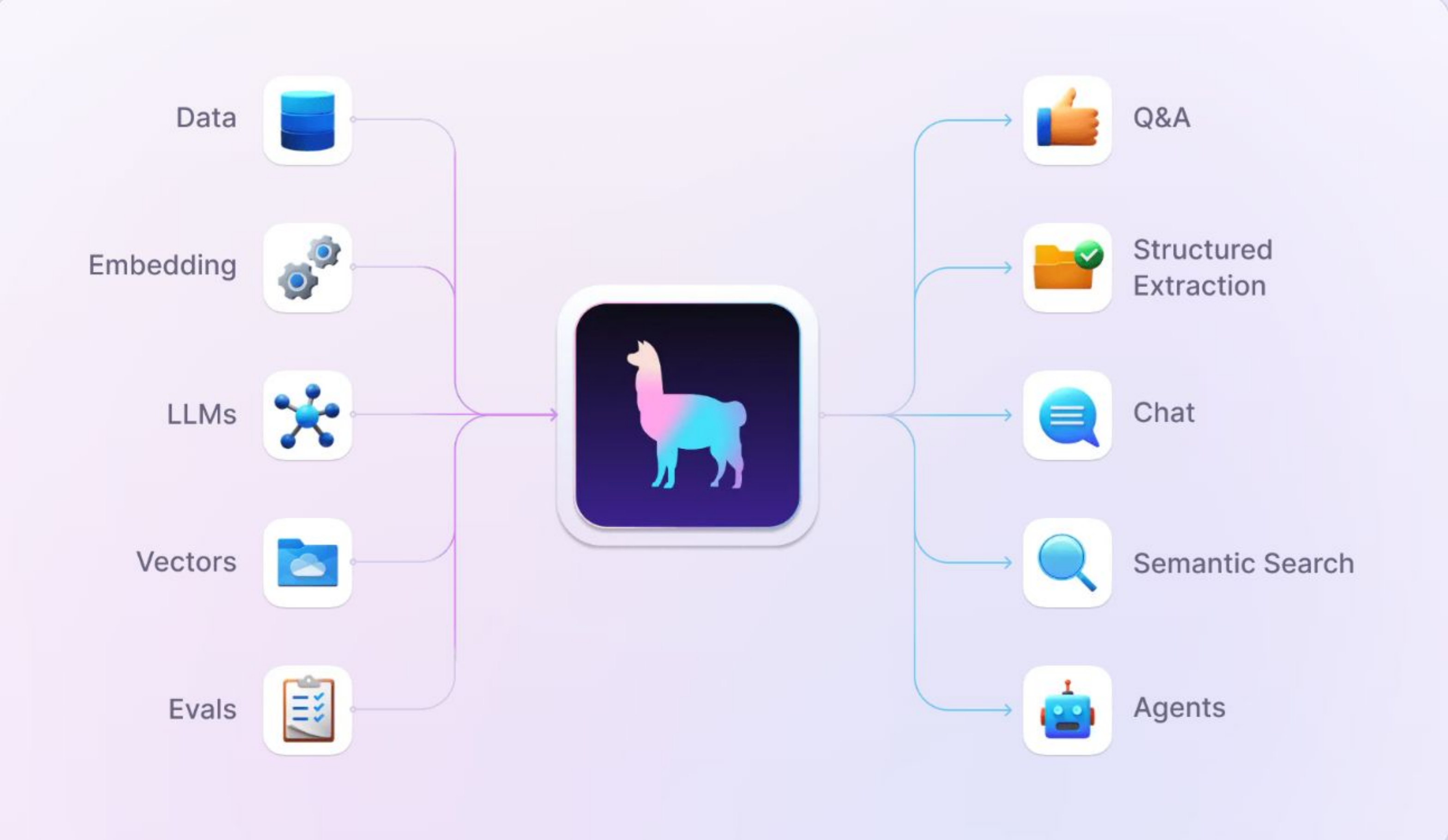
- Core functionality:
  - **Partitioning**: Extraction of structured contents
  - **Cleaning**: Removing unwanted content
  - **Extracting**: Extraction of specific entities within a document
  - **Chunking**: Partition documents into semantic units



```
0
ear Ended December 31,
1
2021                        2020                        2019
2                                                          Revenues
3                                                   Automotive sales
$    44,125                  $    24,604                  $    19,358
4                                      Automotive regulatory credits
1,465                        1,580                         594
5                                                 Automotive leasing
1,642                        1,052                         869
6                                         Total automotive revenues
47,232                       27,236                       20,821
7                                      Energy generation and storage
2,789                        1,994                        1,531
8                                                 Services and other
3,802                        2,306                        2,226
9                                                     Total revenues
53,823                       31,536                       24,578
10                                                  Cost of revenues
11                                                  Automotive sales
32,415                       19,696                       15,939
12                                                Automotive leasing
978                          563                          459
13                               Total automotive cost of revenues
33,393                       20,259                       16,398
14                                 Energy generation and storage
2,918                        1,976                        1,341
15                                              Services and other
3,906                        2,671                        2,770
16                                         Total cost of revenues
40,217                       24,906                       20,509
17                                                    Gross profit
13,606                        6,630                        4,069
18                                              Operating expenses
19                                        Research and development
2,593                        1,491                        1,343
20                            Selling, general and administrative
4,517                        3,145                        2,646
21                                        Restructuring and other
```

Llama Index framework

A framework that combines LLM and data for Q&A and Chat!



How does the LLM know which parts of the document are the most relevant to the query?

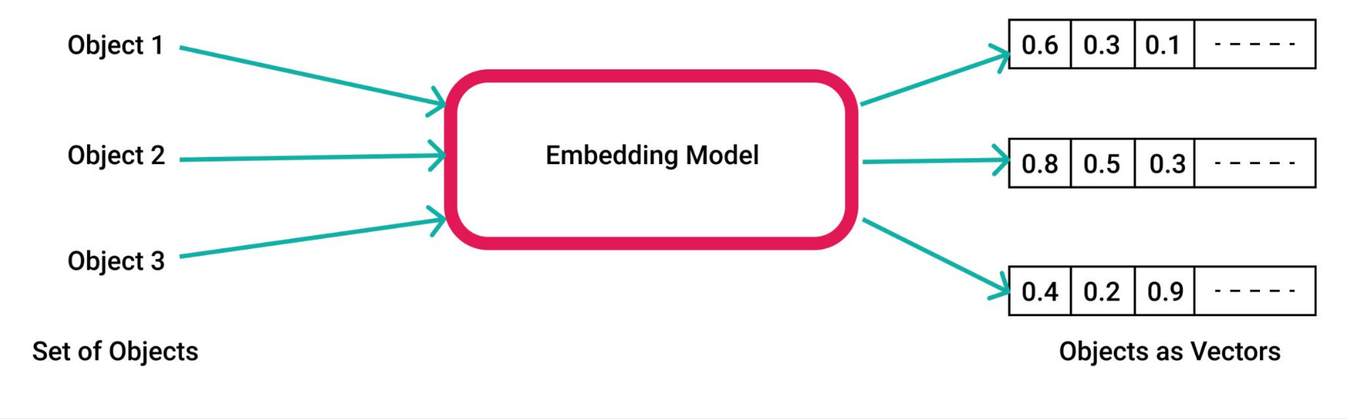
What is data embedding?

- Represent textual data as vector representations in high dimensional space.
- BAAI/bge-large-en-v1.5

- Example:
  - Sentence 1: LLMs are AI models that understand and generate human language.
  - Sentence 2: Large Language Models can comprehend and produce natural language.

- Cosine Similarity: 0.59

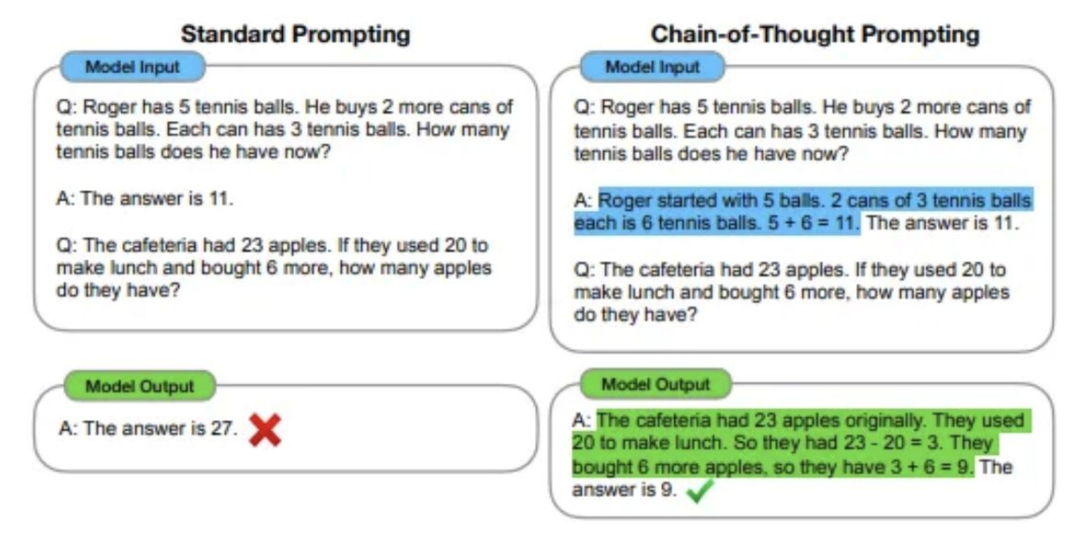
- Similarity ranges from -1 (Least Similar) to +1 (Most similar)



Can quality prompts produce better results?

Chain of Thought Prompting

Chain of thought prompting is a technique used with AI language models to help them solve complex problems. It involves guiding the AI to articulate its reasoning step-by-step, much like how a human would think aloud while solving a problem. By explicitly asking the AI to describe each step in its thought process, this method can improve the AI's accuracy and provide insights into how it arrives at its conclusions. Essentially, it makes the AI "think" in a structured and transparent way, making it easier to follow and verify the logic behind its answers.



Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

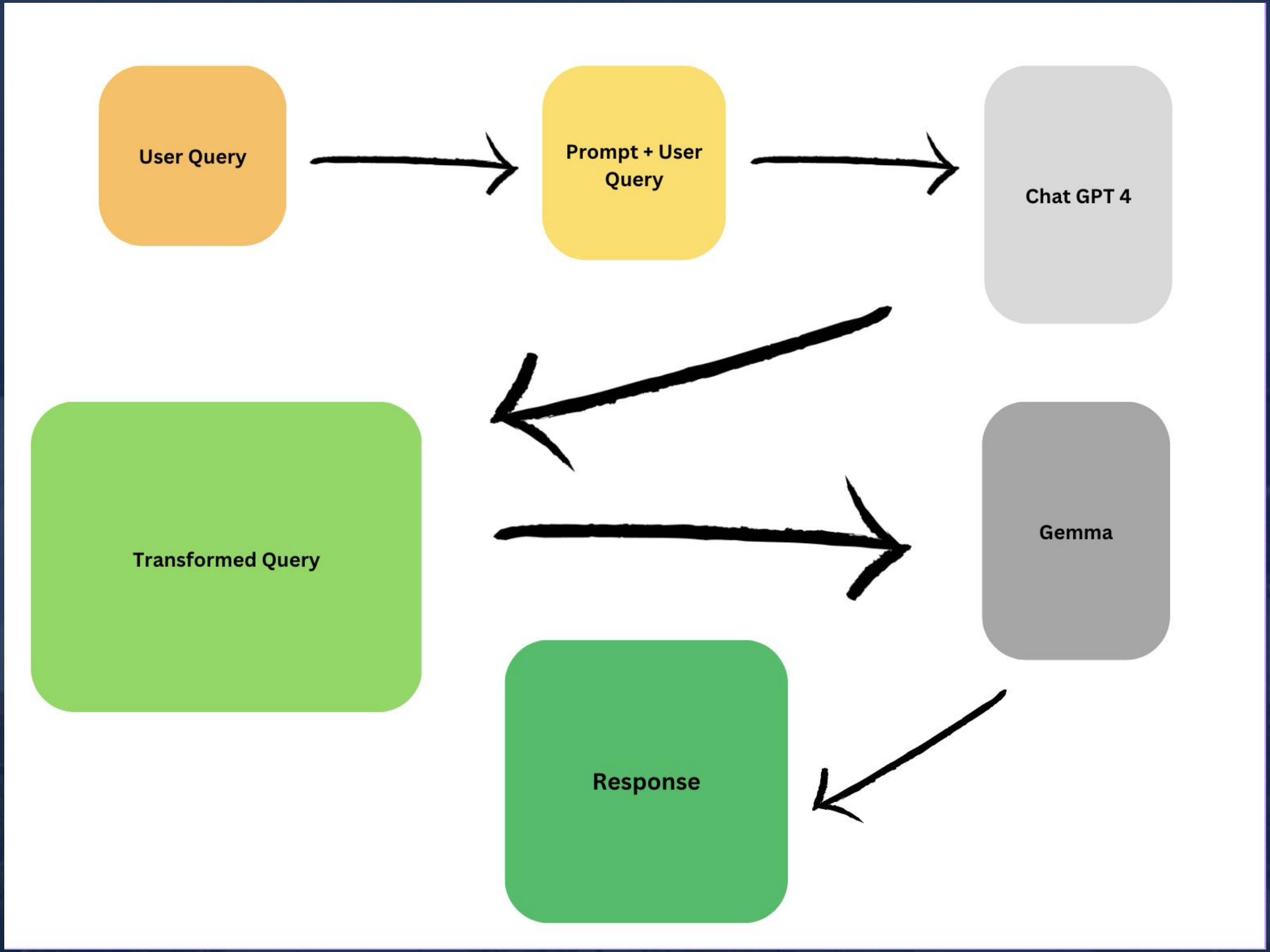
Model Output

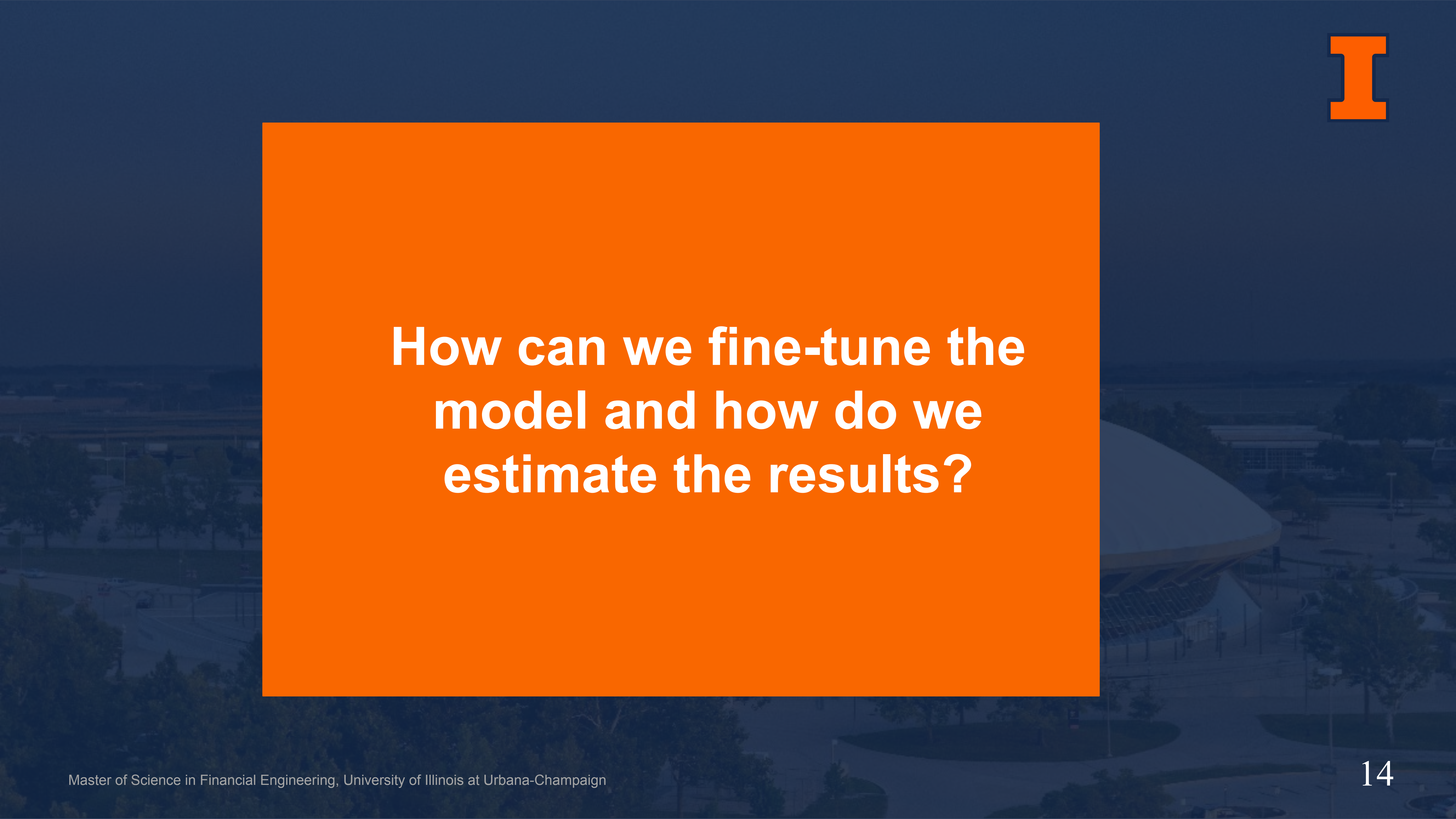
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Chain of Thought Prompting

[CONTEXT] + [SPECIFIC INFORMATION] + [INTENT] + [RESPONSE FORMAT] = PERFECT PROMPT

"I am fine-tuning a Large Language Model to analyze 10-K forms for financial analysis. I want to utilize chain of thought prompting so that my model can produce more accurate answers. Can you refine the following question in a way that would produce a more accurate response from my model? It may be useful to include where the information can be found or how to calculate the answer if calculations are required. The refined question should have the following format: [context] + [specific information] + [intent] + [response format]. The only output should be the refined question. Do not include any additional information, explanation, titles, or headers in the output. The output should be in a normal formatting with no bullet points."

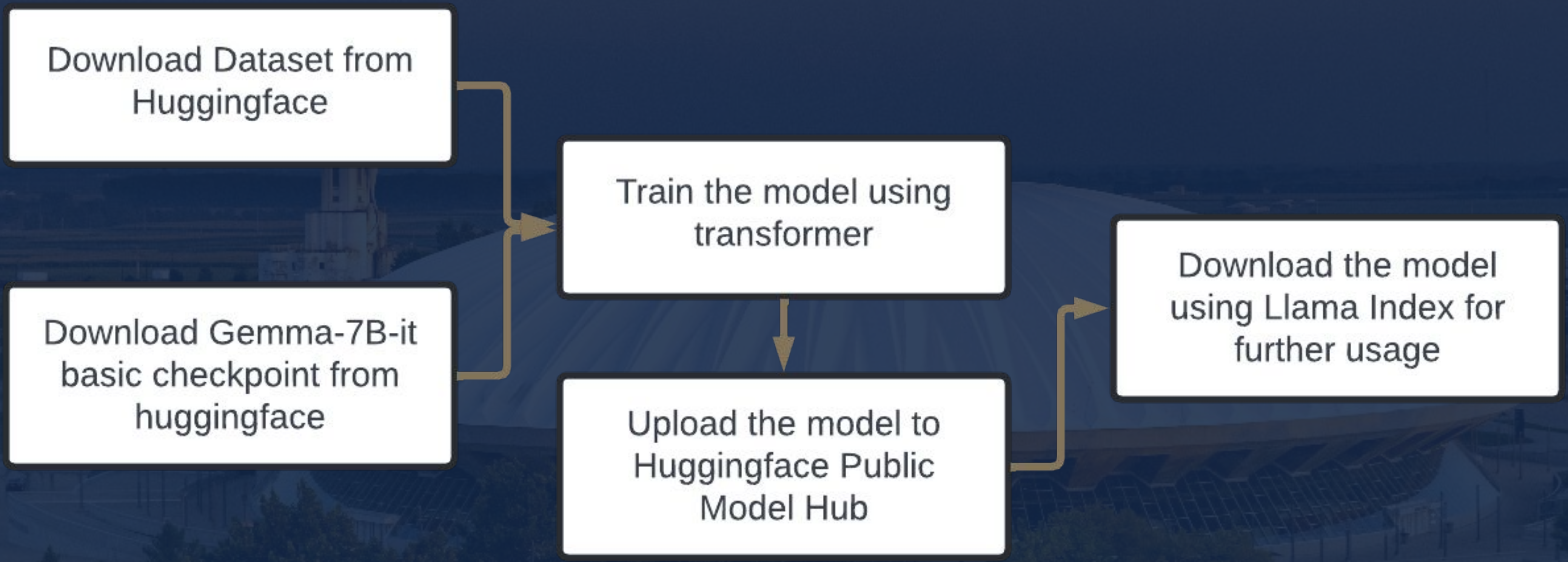




How can we fine-tune the model and how do we estimate the results?

Model Fine-Tuning

Problem: Llama Index does not support Model Fine-tuning!



Download Dataset from Huggingface

Download Gemma-7B-it basic checkpoint from huggingface

Train the model using transformer

Upload the model to Huggingface Public Model Hub

Download the model using Llama Index for further usage

Dataset: 10-k Documents from 1990 to 2020, splitted into chunks

Model Fine-Tuning

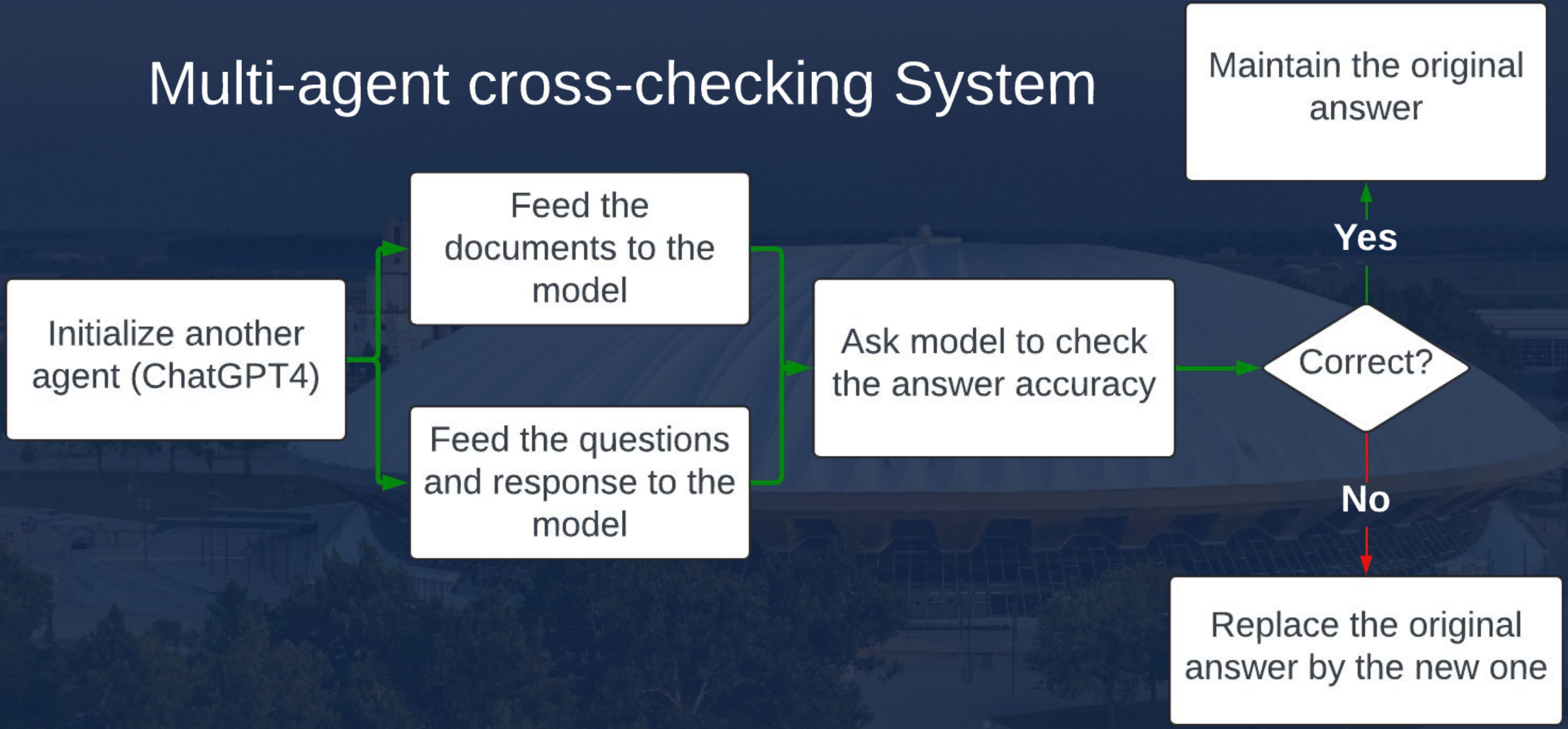
Trainable parameters: 200 Million	Total Parameters: **8.7 billion**	Percentage: 2.29%
Model size: **31.81 GiB**	Hardware: **24GB+ RAM**	8K Context length
Transformer structure: 28 layers	16 attention heads	3072 embedding dimension

Dataset: JanosAudran/financial-reports-sec		
Dataset size: 15 GiB Pure text	Number of rows: 68,382,177	Chunk size in one row: 256

Training parameters		
Batch size: 20	Adjusted Learning rate	Optimizer: Adafactor
maximum steps: 20	Start loss: 10.87	End loss: 0.27

Performance Evaluation

Multi-agent cross-checking System



Initialize another agent (ChatGPT4)

Feed the documents to the model

Feed the questions and response to the model

Ask model to check the answer accuracy

Correct?

Yes

Maintain the original answer

No

Replace the original answer by the new one

Performance Evaluation

"You are a financial analyst.”	Define the role
"Utilizing your expertise, check the accuracy of the question and answer regarding this company.”	
Task	
"Give a score to this answer from 0 to 5, and state the reason.”	Task Breakdown
“If the answer is incorrect, you need to give the correct answer."	Task Breakdown
"Return the score first and use '/' to separate the score and the explanation or correct answer. “	Format

Question	Original response	Level	Agent Response
How much of the business is purchase vs. refinance activity?	Purchase activity accounted for $29.5 billion of the business, while refinance activity amounted to $0 million in the year 2023.	moderate	0/ The answer provided is incorrect. The correct answer is that the information provided does not specify the exact breakdown between purchase and refinance activity for the company.
How does the firm intend to grow its market share, and what was the firm’s market share this period?	The firm intends to grow its market share by developing and retaining experienced teams, maintaining a significant cash balance, actively managing inventory investments, offering new home communities, delivering high-quality homes and customer experiences, investing in desirable markets, …	moderate	4/ The answer accurately describes how the firm intends to grow its market share by outlining various strategies and initiatives. However, the market share value provided in the answer is incorrect. The correct market share value for this period is approximately $32.9 billion based on the closing price of the common stock held by non-affiliates.

Areas of improvement

1. Question classification:
   Different prompt for qualitative / quantitative questions?

2. Model response time (Current time: 30 to 40 seconds)

3. Minimize model output **uncertainty**
   Keep the answer **stable**!

Summary

- Data Preprocessing
  - Unstructured.io
- Data Embedding
  - BAAI embedding
- LLM
  - Gemma (Pre-trained LLM by Google)
  - Fine Tuning the model using HuggingFace dataset
- Chain of thought prompting
  - Use of another model to generate prompts
- Multi-agent cross checking system for performance evaluation

Appendices

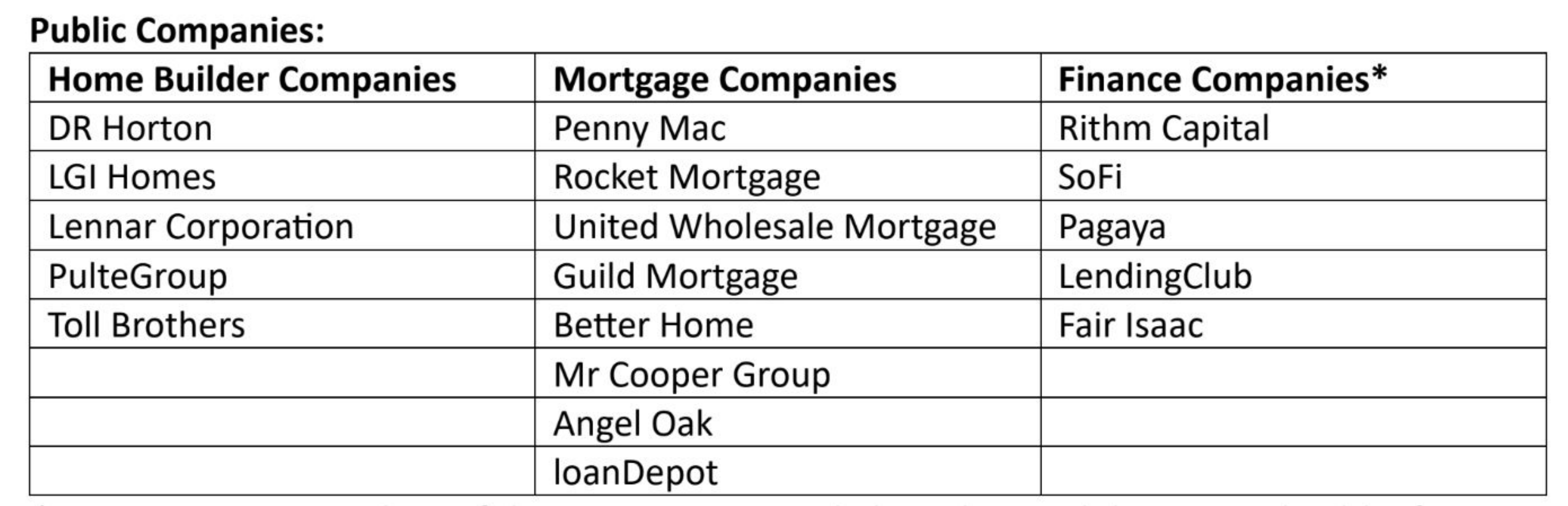
What are 10K forms?

- The data we are working with is a collection of the past 5 years of 10K forms for each of these 18 companies, totaling 90 documents
- 10-K forms are comprehensive annual reports filed by publicly traded companies in the United States, detailing their financial performance, operations, corporate governance, and risk factors as required by the Securities and Exchange Commission (SEC).
- Data types within 10-K forms include quantitative financial information such as balance sheets, income statements, cash flow statements, as well as qualitative disclosures about business operations, risk factors, and management's discussion and analysis of financial conditions.

Why is Preprocessing important?

- The data needs to be transformed into a format that can be more easily digested and evaluated by our LLM.
- This will improve the accuracy of the model and retrieval of information.
- Preprocessing often involves tokenization, which is splitting the document into more manageable pieces or chunks.
- These tokens are then combined with various metadata, such as the name and year of the corresponding document or summaries, to create identifiable nodes of data that can be more easily retrieved.
- 10K forms contain various tabular and graphical information, and it is important to transform this information into something the LLM can easily process.

What data are we collecting?

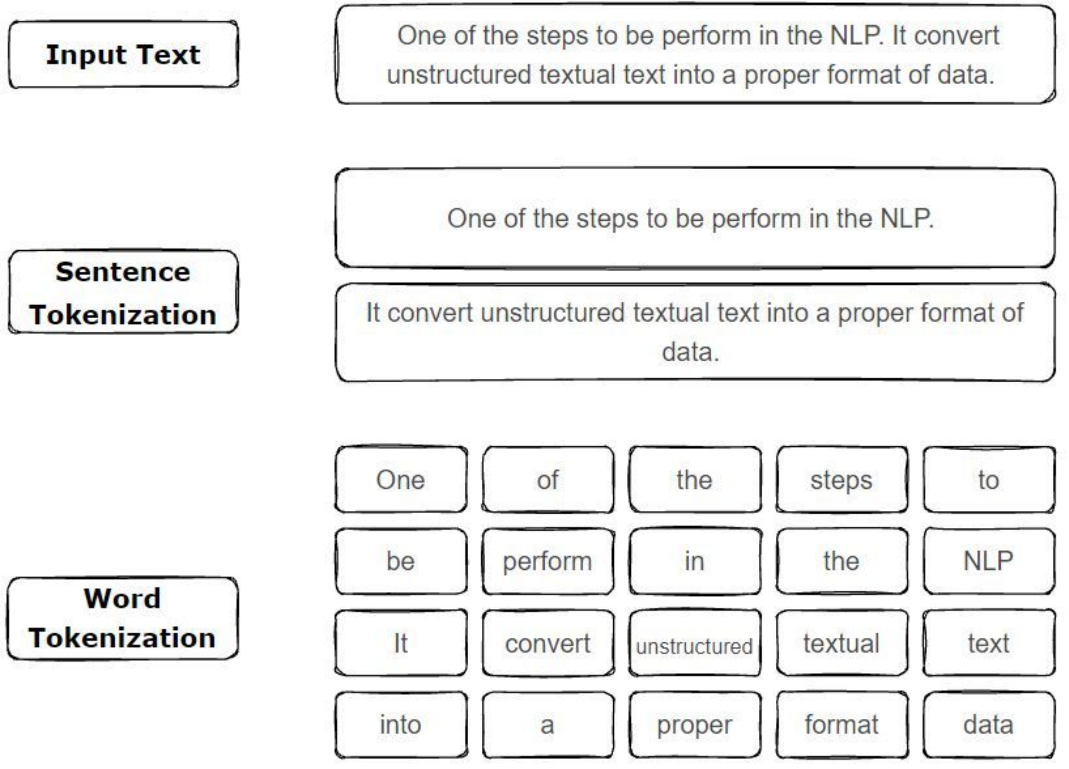


Public Companies:

Home Builder Companies	Mortgage Companies	Finance Companies*
DR Horton	Penny Mac	Rithm Capital
LGI Homes	Rocket Mortgage	SoFi
Lennar Corporation	United Wholesale Mortgage	Pagaya
PulteGroup	Guild Mortgage	LendingClub
Toll Brothers	Better Home	Fair Isaac
	Mr Cooper Group	
	Angel Oak	
	loanDepot	

- Listed above are 18 companies whose financial statements contain important information for JIA Finance to make underwriting decisions
- These companies consist of home builders companies, mortgages companies, and finance companies

Tokenization



Input Text

One of the steps to be perform in the NLP. It convert unstructured textual text into a proper format of data.

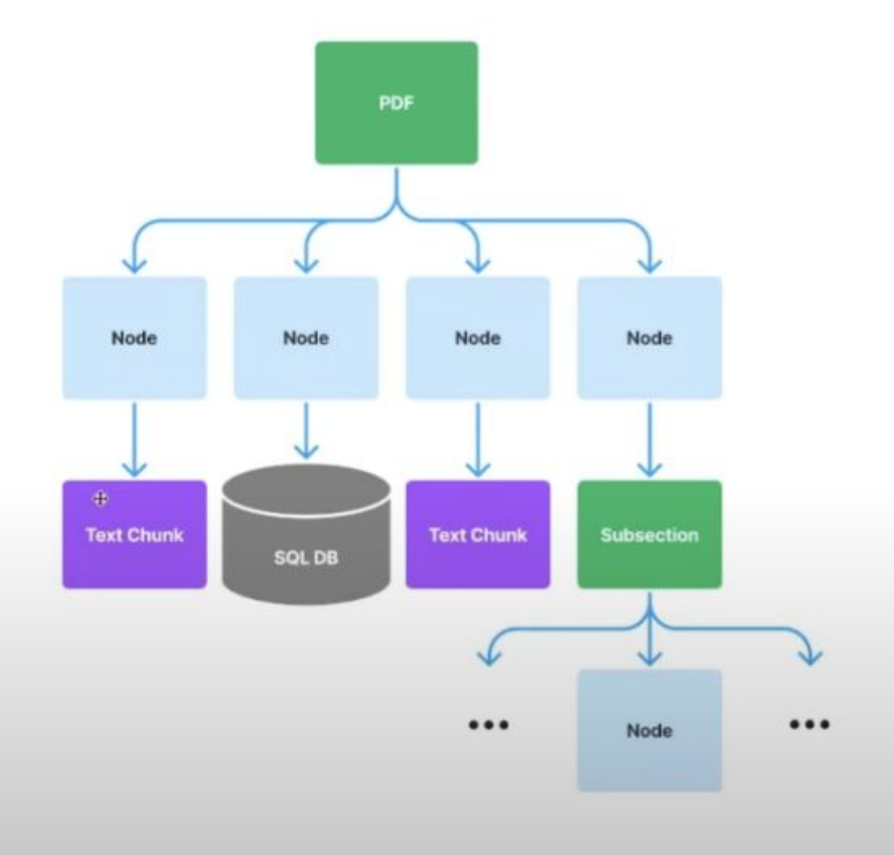
Sentence Tokenization

One of the steps to be perform in the NLP.

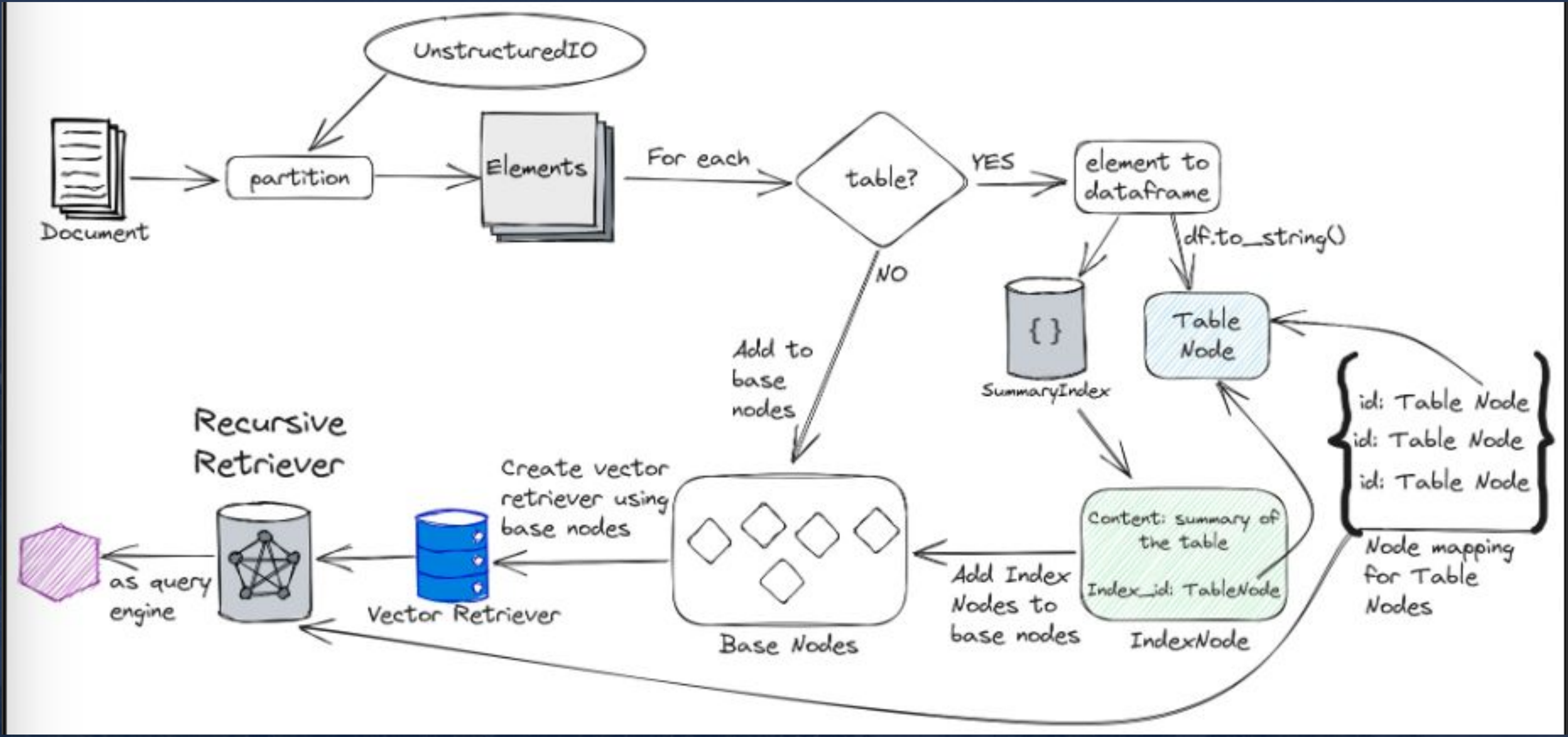
It convert unstructured textual text into a proper format of data.

Word Tokenization

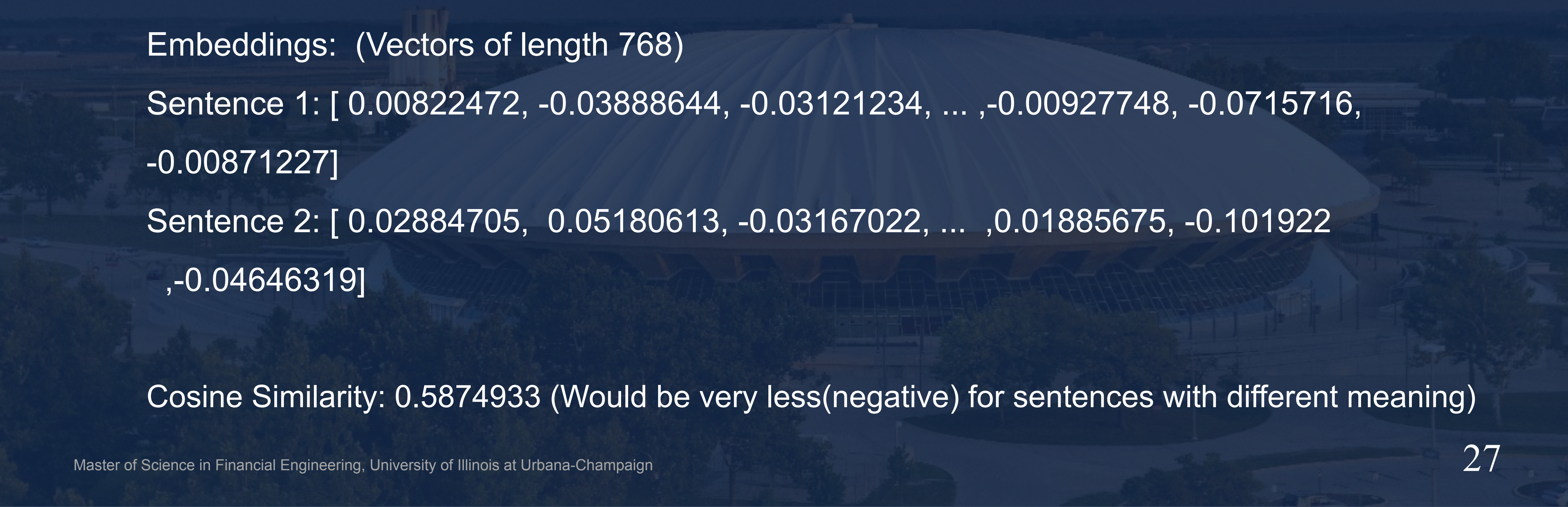
One	of	the	steps	to
be	perform	in	the	NLP
It	convert	unstructured	textual	text
into	a	proper	format	data



Recursive RAG Retriever



Embeddings



Sentence 1: LLMs are AI models that understand and generate human language.

Sentence 2: Large Language Models can comprehend and produce natural language.

Embeddings: (Vectors of length 768)

Sentence 1: [0.00822472, -0.03888644, -0.03121234, ... ,-0.00927748, -0.0715716, -0.00871227]

Sentence 2: [0.02884705, 0.05180613, -0.03167022, ... ,0.01885675, -0.101922 ,-0.04646319]

Cosine Similarity: 0.5874933 (Would be very less(negative) for sentences with different meaning)