

Large Language Models in Education

Group 08

Group Members

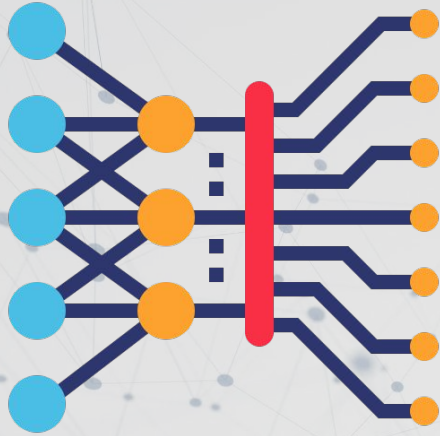
- E/17/297 Rupasinghe T. T. V. N.
- E/17/206 Manohara H. T,
- E/17/148 Kalpana M. W. V.

Supervisors

- Dr. Damayanthi Herath
- Prof. Roshan Ragel
- Dr. Isuru Nawinne
- Dr. Shamane Sri



Introduction



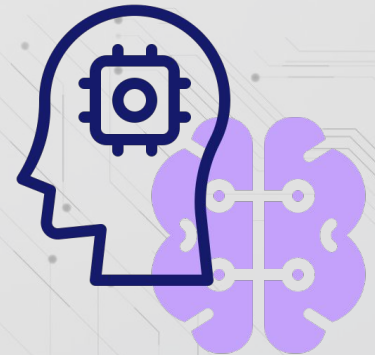
- **Deep learning -> AI type**

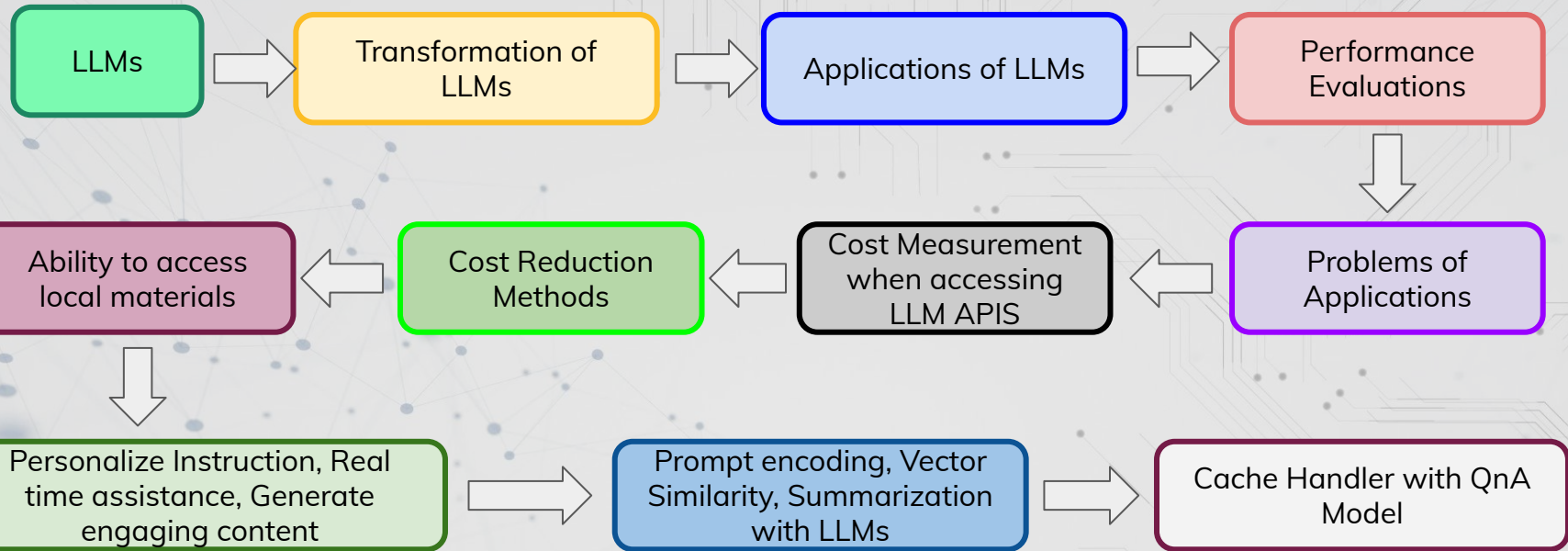
- **Generative AI**

- Makes new data using generative models
 - Take input and learn the patterns by training and then generate new data with same characteristics.

- **LLMs**

- Text generating part of Generative AI
 - Form of Generative AI
 - Data + Architecture + Training
 - Prototype language applications incredibly fast
 - Transformer models





Research on LLMs in Education



Cost Reduction Methods

- Prompt Adaptation
- LLM Approximation
- LLM Cascade

Prompt Adaptation

Prompt: Q1+A1, Q2+A2, Q3+A3, Q4+A4

Q: What is the result of N₂ and O₂ at high temperature?

Prompt: Q1+A1, Q2+A2, Q3+A3, Q4+A4

Q: What helps prey hide?

Query Concatenator

Prompt: Q1+Q2, A1+A2, Q3+Q4, A3+A4

Q: What is the result of N₂ and O₂ at high temperature?
Q2: What helps prey hide?

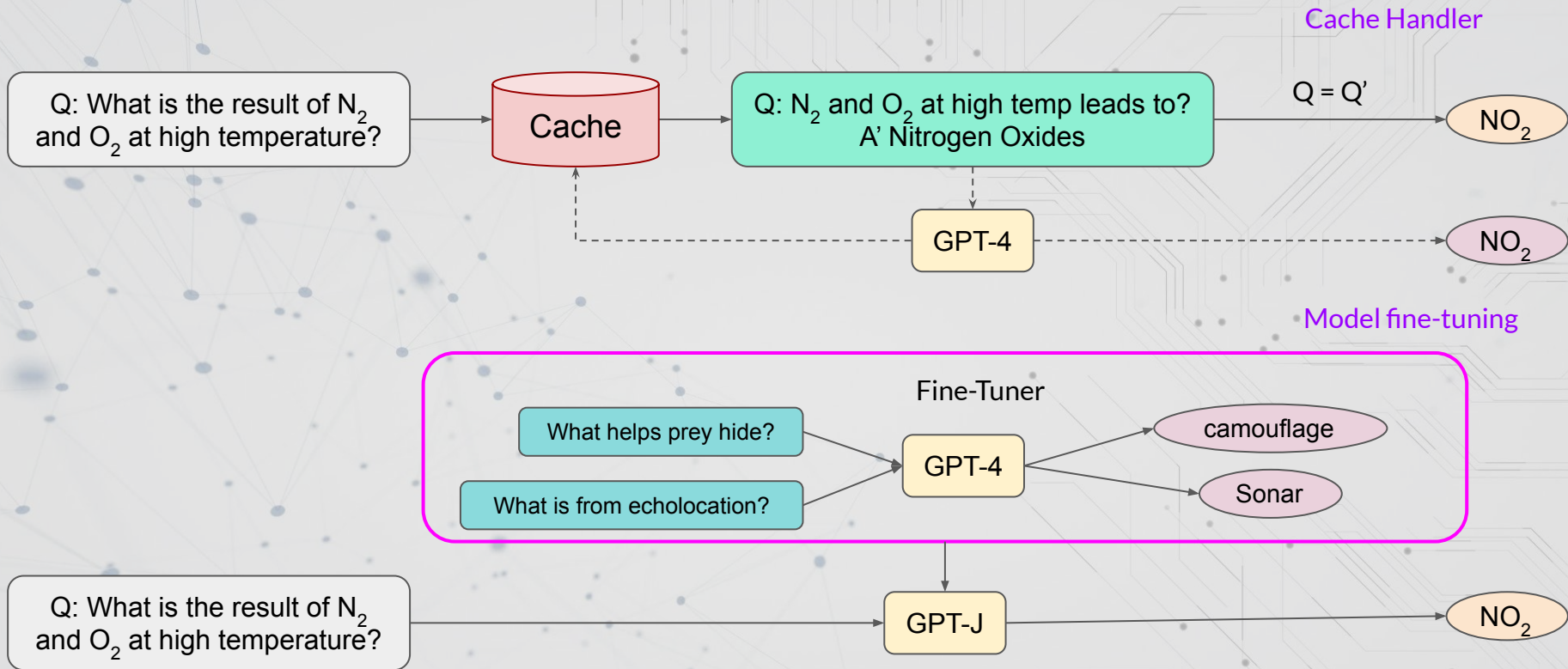
GPT-4

NO₂

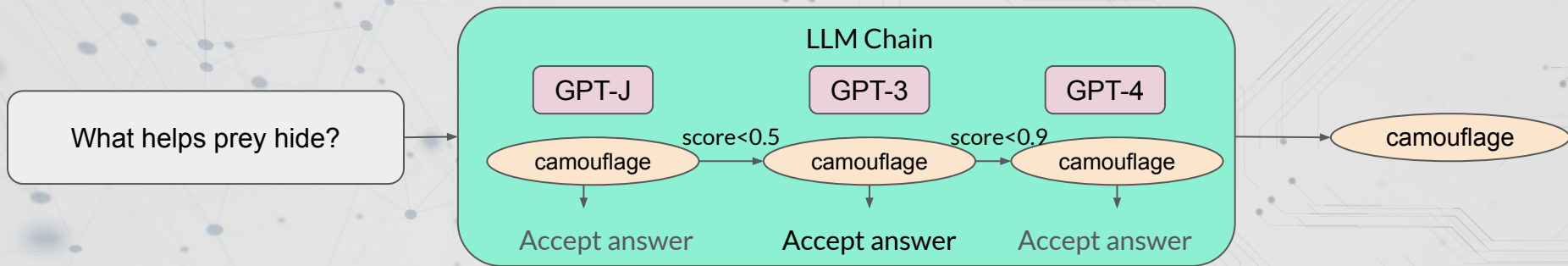
camouflage

Query Concatenation

LLM Approximation



LLM Cascade



Conclusion

- Cache Handler was selected as the cost reduction method based on,
 - Reduction of API calls

Cost Measurement

- Cost of LLM APIs based on
 - Number of input **Tokens** (unit of text)
 - Number of output **Tokens**
 - Fixed cost per Request

Summary of commercial LLM APIs

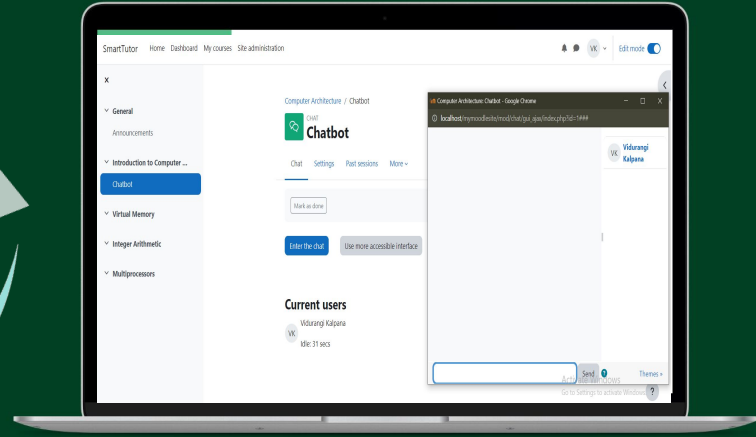
Provider	API	Size/B	Cost (USD)		
			10M input tokens	10M output tokens	request
OpenAI	GPT-Curie	6.7	2	2	0
	ChatGPT	NA	2	2	0
	GPT-3	175	20	20	0
	GPT-4	NA	30	60	0
AI21	J1-Large	7.5	0	30	0.0003
	J1-Grande	17	0	80	0.0008
	J1-Jumbo	178	0	250	0.005
Cohere	Xlarge	52	10	10	0
ForeFront AI	QA	16	5.8	5.8	0
Textsynth	GPT-J	6	0.2	5	0
	FAIRSEQ	13	0.6	15	0
	GPT-Neox	20	1.4	35	0

High cost of LLM APIs



When accessing LLMs through API calls, APIs get high cost

Cost Effective Intelligent Tutor



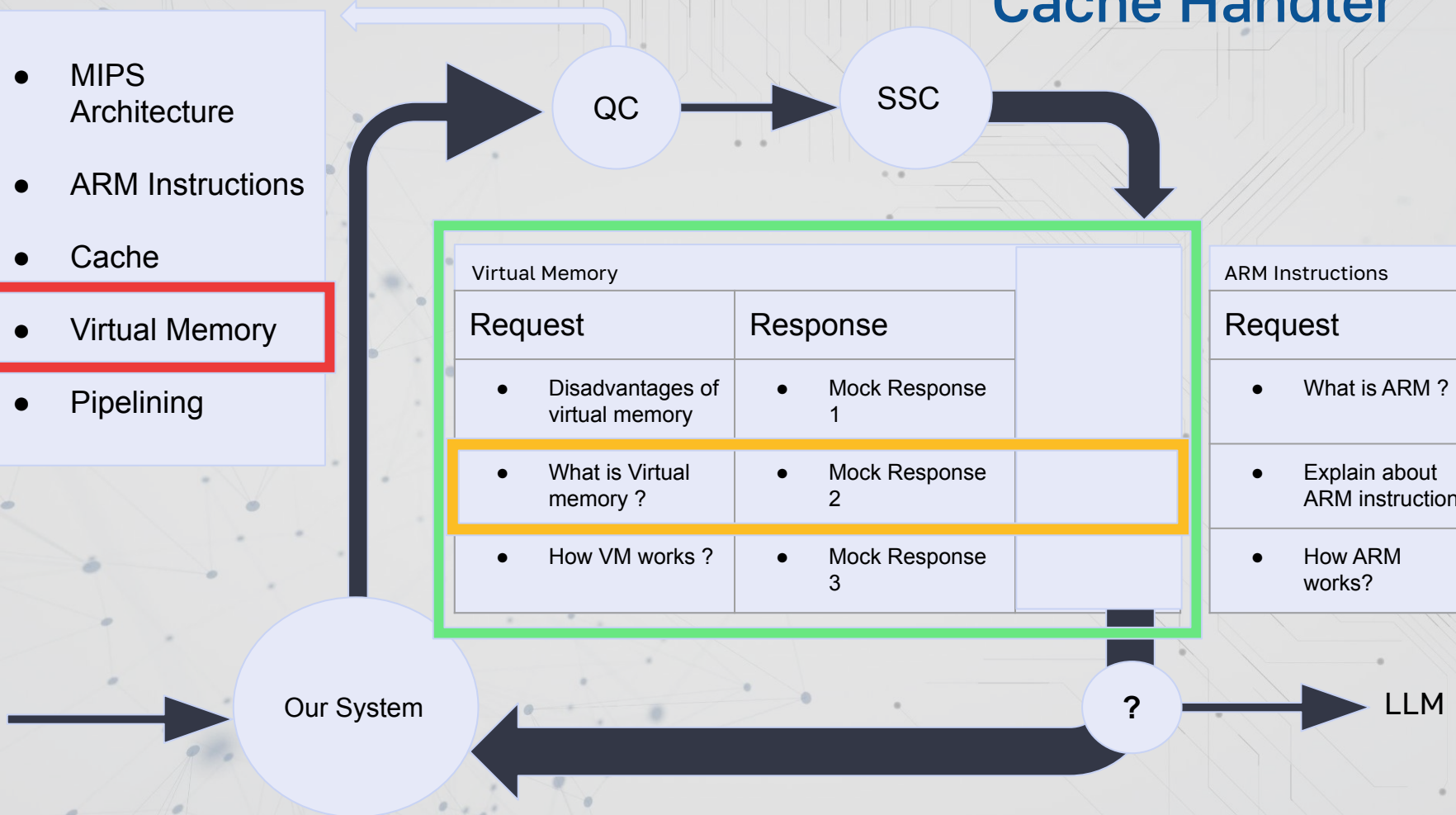
- Cost Reduction using cache
 - Storing the response locally in a cache when submitting a query to an LLM API & verifying a similar question has been previously answered
- Ability to access local materials without accessing external LLMs

Cache Handler

- MIPS Architecture
- ARM Instructions
- Cache
- Virtual Memory
- Pipelining

Virtual Memory		
Request	Response	
<ul style="list-style-type: none">• Disadvantages of virtual memory	<ul style="list-style-type: none">• Mock Response 1	
<ul style="list-style-type: none">• What is Virtual memory ?	<ul style="list-style-type: none">• Mock Response 2	
<ul style="list-style-type: none">• How VM works ?	<ul style="list-style-type: none">• Mock Response 3	

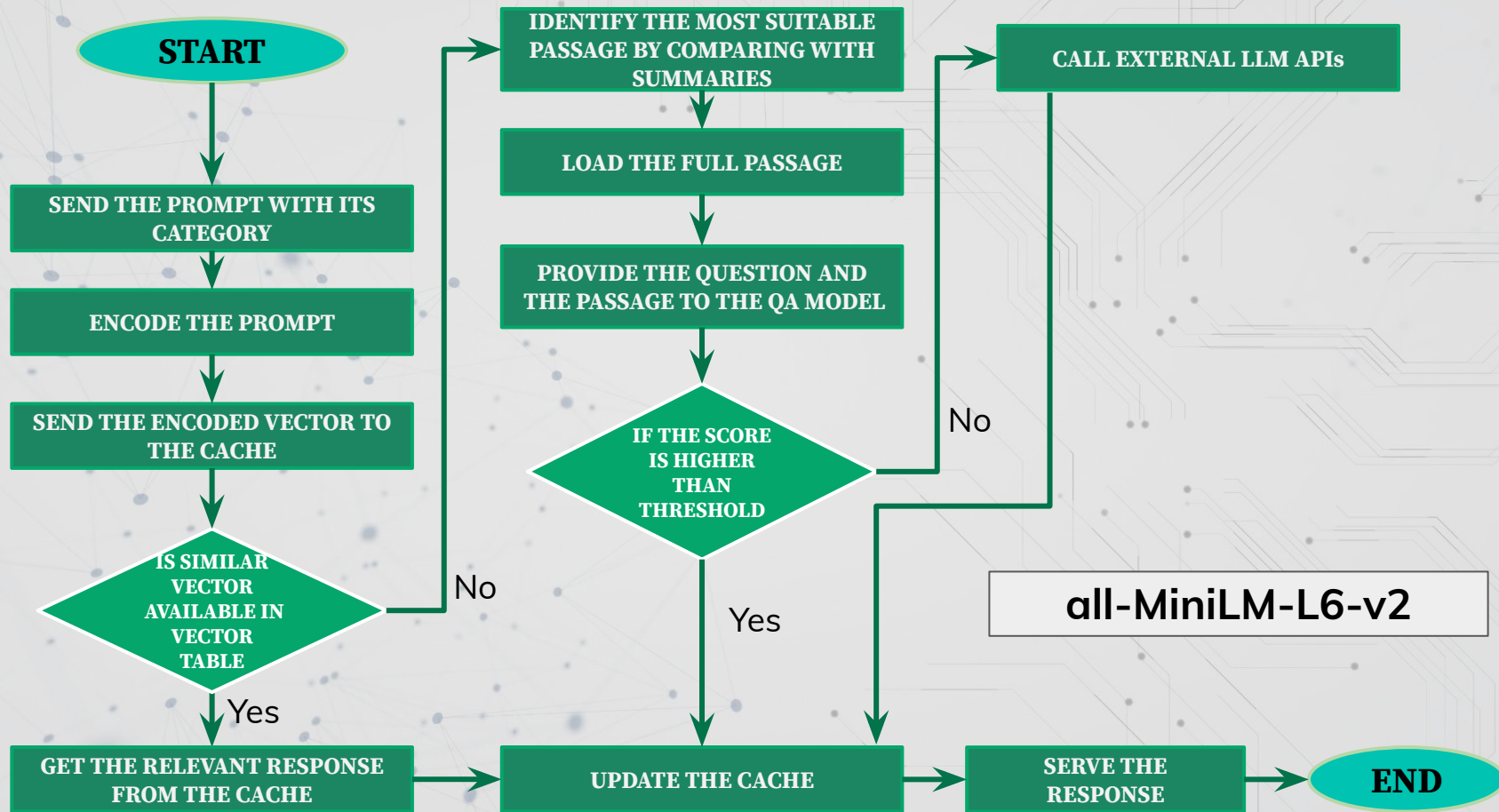
ARM Instructions		
Request	Response	
<ul style="list-style-type: none">• What is ARM ?		
<ul style="list-style-type: none">• Explain about ARM instructions		
<ul style="list-style-type: none">• How ARM works?		



Current Progress



Data Flow of Our System



QA Model Implementation

- Tested with pre-built question answering models.

question-answering (0.06)

deepset/roberta-base-squad2 (0.26)

bert-large-uncased-whole-word-masking-finetuned-squad (0.02)

twmkn9/bert-base-uncased-squad2 (0.70)

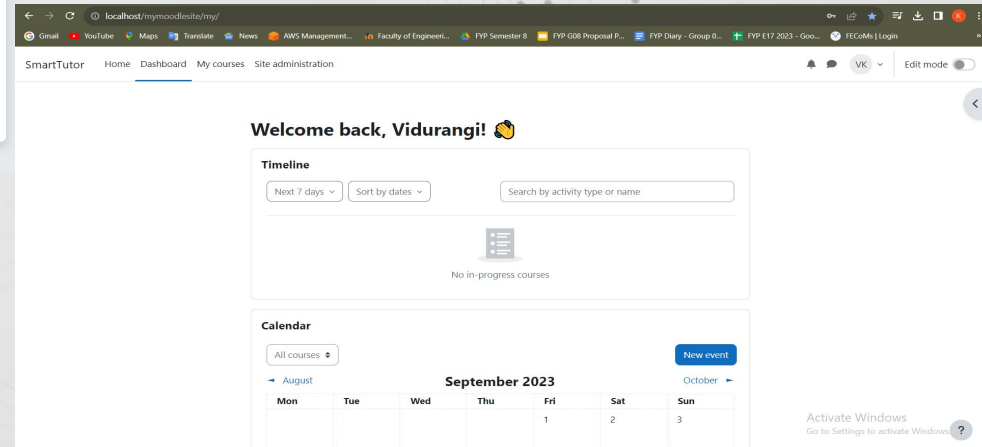
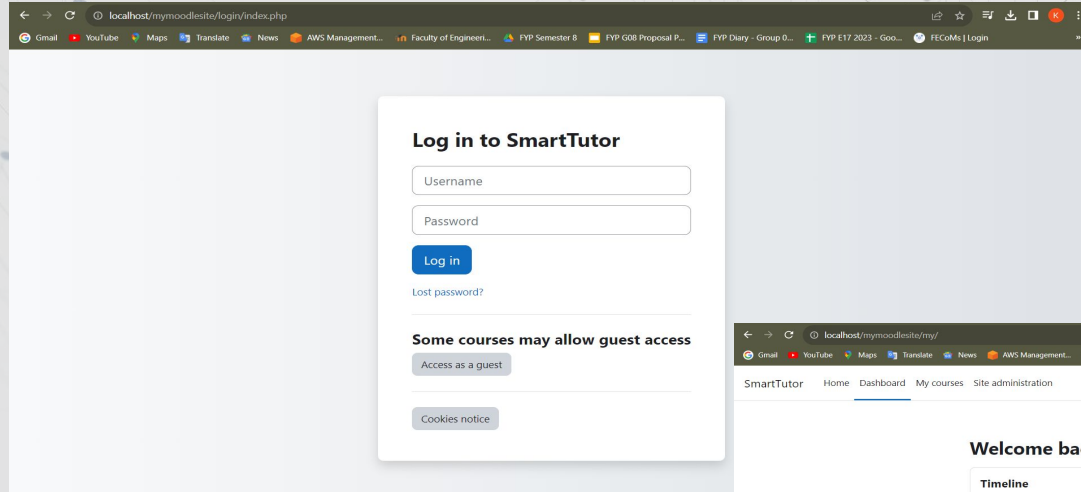
QA Model Implementation cont.

- Creating custom QA models.
 - Focus on BERT models
- Created the Dataset.
- Trained models for our context
 - **bert-base-cased**
 - **electra-base-discriminator**

```
Running Prediction: 100% ██████████ 1/1 [00:00<00:00, 16.63it/s]
[{'id': '00001', 'answer': ['empty']}]
```

```
Some weights of the model checkpoint at twmkn9/bert-base-uncased-squad2 were not used when initializing BertForQuestionAnswering: ['bert.pooler.dense.bias']
- This IS expected if you are initializing BertForQuestionAnswering from the checkpoint of a model trained on another task or with another architecture (e.g. another pre-training objective)
- This IS NOT expected if you are initializing BertForQuestionAnswering from the checkpoint of a model that you expect to be exactly identical (initializing from a pre-trained model)
{'score': 0.32008060812950134, 'start': 1282, 'end': 1362, 'answer': 'both identifies the type of interrupt, and provides the handler at the same time'}
```

Implementation of the Interface



Moodle Frontend View

The screenshot displays the Moodle Frontend View for a chatbot interface. The browser address bar shows the URL `localhost/mymoodle/site/mod/chat/view.php?id=2&forceview=1`. The page title is "SmartTutor" and the navigation menu includes "Home", "Dashboard", "My courses", and "Site administration".

The main content area is titled "Computer Architecture / Chatbot" and features a green chatbot icon. Below the icon are navigation links for "Chat", "Settings", "Past sessions", and "More". A "Mark as done" button is visible. Two buttons are present: "Enter the chat" (highlighted in blue) and "Use more accessible interface".

The "Current users" section lists "Vidurangi Kalpana" with the initials "VK" and a status of "Idle: 31 secs".

An inset window titled "Computer Architecture: Chatbot - Google Chrome" shows a chat interface with a user profile for "Vidurangi Kalpana" (VK) and a "Send" button. The chat area is currently empty.

Cache Implementation

- Implemented using Least frequency used (LFU) policy.

```
] test_sentence = "can you please tell me how system performance is affected by s...  
category = "vm"  
response_for_test_sentence = give_the_response_v2(test_sentence, category)  
print(response_for_test_sentence)
```

```
***** From Cache *****
```

```
2  
Resp 23
```

```
] print(cacheVM.cache_df)
```

	Question	Response	Access Count
0	How does TLB caching improve virtual memory pe...	Resp 23	4
1	What is the purpose of the Translation Lookasi...	Resp 22	3
2	How does swapping affect system performance?	Resp 17	2
3			0

```
test_sentence = "can you please tell me how virtual memory help in managing me...  
category = "vm"  
response_for_test_sentence = give_the_response_v2(test_sentence, category)  
print(response_for_test_sentence)
```

```
***** From Cache *****
```

```
3  
Resp 9
```

```
print(cacheVM.cache_df)
```

	Question	Response	Access Count
0	How does swapping affect system performance?	Resp 17	4
1	How does TLB caching improve virtual memory pe...	Resp 23	15
2	What is the purpose of the Translation Lookasi...	Resp 22	1
3	How does virtual memory help in managing memor...	Resp 9	4

```
: test_sentence = "can you please tell me how virtual memory support memory prote...  
category = "vm"  
response_for_test_sentence = give_the_response_v2(test_sentence, category)  
print(response_for_test_sentence)
```

```
***** Calling API *****
```

```
Least accessed question: Question      What is the purpose of the Translation  
Response                               Resp 22  
Access Count                           1  
Name: 2, dtype: object  
Removing record with index 2  
API response for -> can you please tell me how virtual memory support memory prot...  
processes?
```

```
: print(cacheVM.cache_df)
```

	Question	Response	Access Count
0	How does swapping affect system performance?	Resp 17	4
1	How does TLB caching improve virtual memory pe...	Resp 23	15
2	What is the purpose of the Translation Lookasi...	Resp 22	1
3	How does virtual memory help in managing memor...	Resp 9	4

```
test_sentence = "can you please tell me how virtual memory support memory prot...  
category = "vm"  
response_for_test_sentence = give_the_response_v2(test_sentence, category)  
print(response_for_test_sentence)
```

```
***** From Cache *****
```

```
2  
Resp 15
```

```
print(cacheVM.cache_df)
```

	Question	Response	Access Count
0	How does swapping affect system performance?	Resp 17	4
1	How does TLB caching improve virtual memory pe...	Resp 23	15
2	How does virtual memory support memory protect...	Resp 15	2
3	How does virtual memory help in managing memor...	Resp 9	4

Problems encountered during the proposed study

- The accuracy is not considerable in the custom QA models.
- Moodle Instance Installation - Server plugins were not installed correctly.
- Selecting a suitable Cache eviction policy.

Installation - Moodle 4.2.2 (Build: 20230814)

Moodle 4.2.2 (Build: 20230814)

For information about this version of Moodle, please see the online [Release Notes](#)

Server checks

Name	Information	Report	Plugin	Status
database	mariadb (10.4.28-MariaDB)	version 10.6.7 is required and you are running 10.4.28		Check
php_setting	opcache.enable	PHP setting should be changed PHP opcode caching improves performance and lowers memory requirements, OPcache extension is recommended and fully supported.		Check
unicode		must be installed and enabled		Pass
php		version 8.0.0 is required and you are running 8.1.1		Pass
pcreunicode		should be installed and enabled for best result		Pass
php_extension	iconv	must be installed and enabled		Pass
php_extension	mbstring	must be installed and enabled		Pass
php_extension	curl	must be installed and enabled		Pass
php_extension	openssl	must be installed and enabled		Pass

How the challenges were handled

- Researched on articles for selecting the appropriate Cache Policy.
- Troubleshooted server plugins by editing the php file in xampp server

Work Plan









Cost Measurement





- Cost of LLM APIs based on
 - Number of input **Tokens** (unit of text)
 - Number of output **Tokens**
 - Fixed cost per Request

What we plan to do

- Select an API
- Measure the Cost of LLM API calls with cache
- Compare and contrast the cost with the accuracy of the outputs with data of already implemented platforms

Model	Input	Output
gpt-3.5-turbo-1106	\$0.0010 / 1K tokens	\$0.0020 / 1K tokens
gpt-3.5-turbo-instruct	\$0.0015 / 1K tokens	\$0.0020 / 1K tokens

Task	Sep 18 - Sep 22	Sep 25 - Sep 29	Oct 02 - Oct 06	Oct 09 - Oct 13	Oct 16 - Oct 20	Oct 23 - Oct 27	Oct 30 - Nov 03	Nov 06 - Nov 10	Nov 13 - Nov 17
Implementation of Custom Models									
Implementation of Cache									
Integrate Cache with Custom QnA models									
Integrate Moodle Instance with Complete Backend Implementation									
Test Frontend with Backend using API calls									
Cost Measurement of API calls									

Task	Sep 18 - Sep 22	Sep 25 - Sep 29	Oct 02 - Oct 06	Oct 09 - Oct 13	Oct 16 - Oct 20	Oct 23 - Oct 27	Oct 30 - Nov 03	Nov 06 - Nov 10	Nov 13 - Nov 17
Compare and Contrast Cost measured from our solution with other solutions' cost									
Finalize Performance Evaluation with cost measurement									
Complete Research Paper (Introduction + Literature + Methodology + Experiments + Results & discussion + Conclusion)									
Complete Documentation (Web Page & Repository)									

A photograph of a classroom with a stack of books on a desk and a blue bag on a chair. The image is overlaid with a green circuit pattern. The text "Thank you.." is centered in the upper right quadrant, flanked by two horizontal green bars.

Thank you..

Q & A

