DEPARTMENT OF COMPUTER ENGINEERING

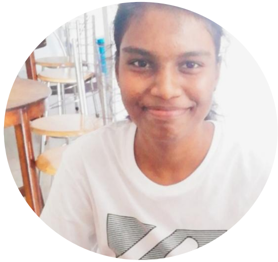
FYP PROJECT | FINAL EVALUATION

GROUP 08

Large Language Models in Education

Our Team

Vishva Nawanjana
E/17/297



Vidurangi Kalpana
E/17/148

Thisara Manohara
E/17/206

Our Supervisors

Dr Damayanthi Herath

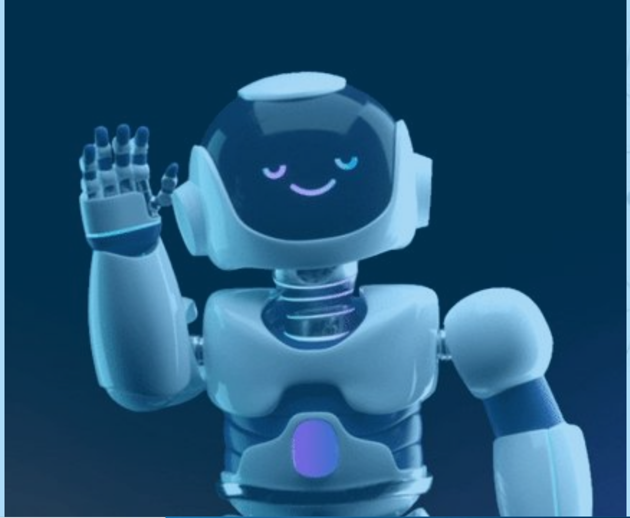
Prof Roshan Ragel

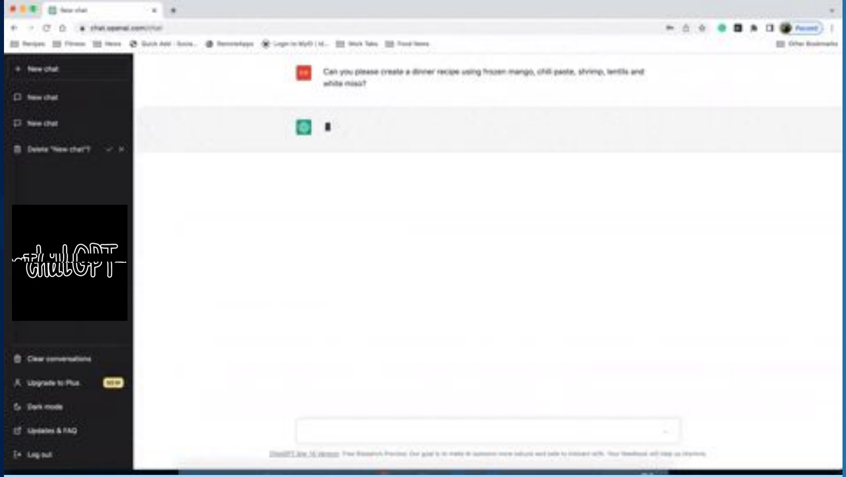
Dr Isuru Nawinne

Dr Shamane Siriwardhana

1.

Background to the Problem

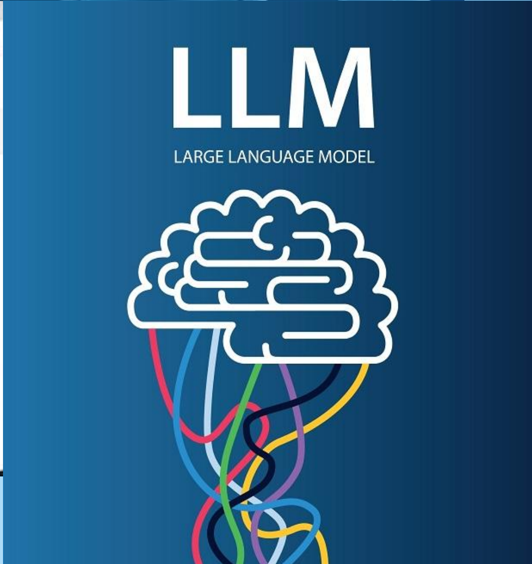




Can you please create a dinner recipe using frozen mango, chili paste, shrimp, lentils and white miso?

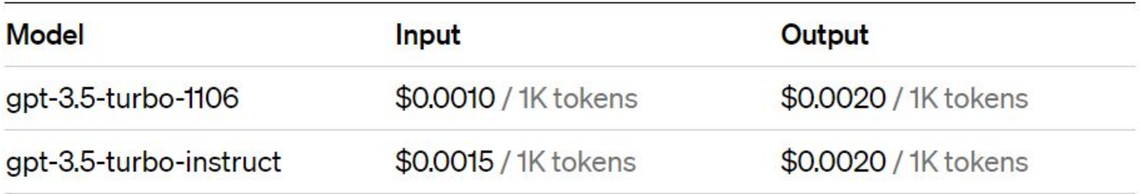
ChatGPT Jan 16 Version. Free Research Preview. Our goal is to make AI systems more natural and safe to interact with. Your feedback will help us improve.

ML Models that are really good at understanding & generating human language based on transformers, a type NN architecture



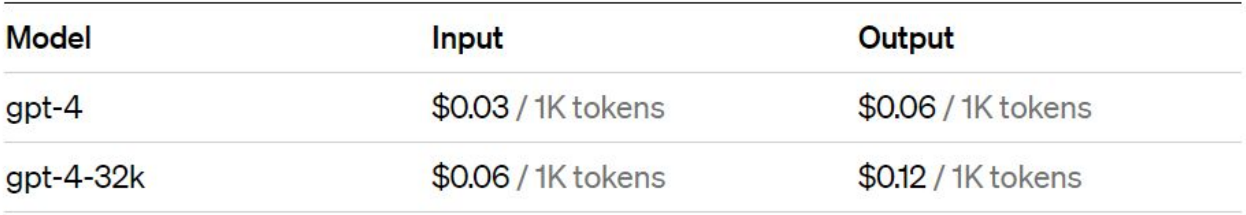
- **High Cost of accessing LLM APIs**
  - Based on the usage volume, measured in terms of API calls or tokens processed.
  - The more API calls or tokens used, the higher the associated cost.

GPT 3.5



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

GPT 4

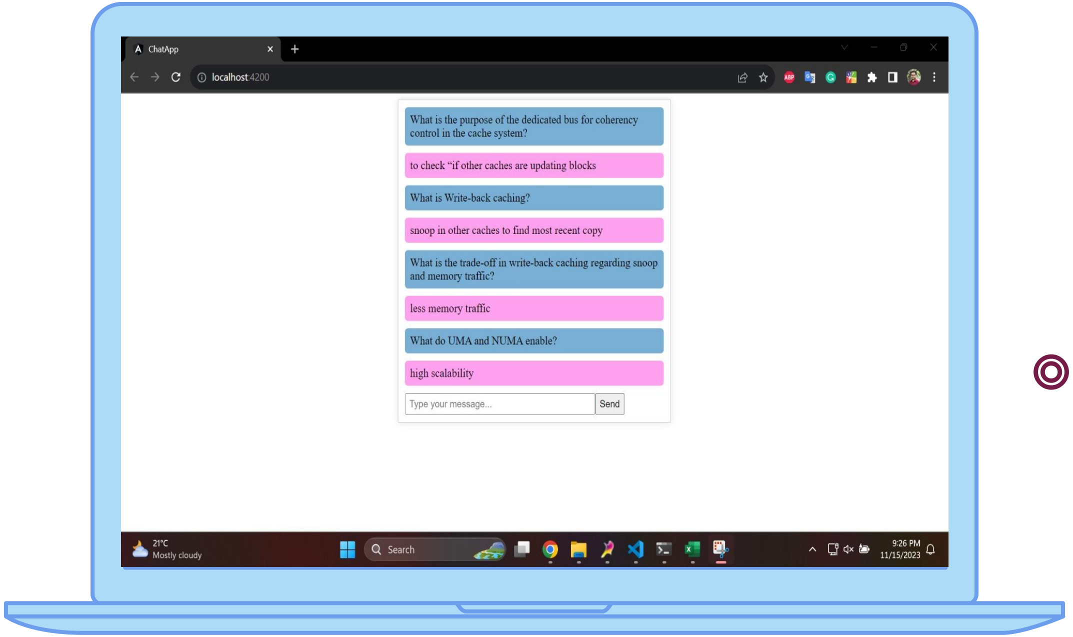


Model	Input	Output
gpt-4	$0.03 / 1K tokens	$0.06 / 1K tokens
gpt-4-32k	$0.06 / 1K tokens	$0.12 / 1K tokens

2.

SOLUTION

Cost Effective Intelligent Tutor



What is the purpose of the dedicated bus for coherency control in the cache system?

to check "if other caches are updating blocks

What is Write-back caching?

snoop in other caches to find most recent copy

What is the trade-off in write-back caching regarding snoop and memory traffic?

less memory traffic

What do UMA and NUMA enable?

high scalability

Type your message... Send

Related Works

- Cost Reduction Methods
- Cost Measurement when accessing LLM APIs

What we implemented

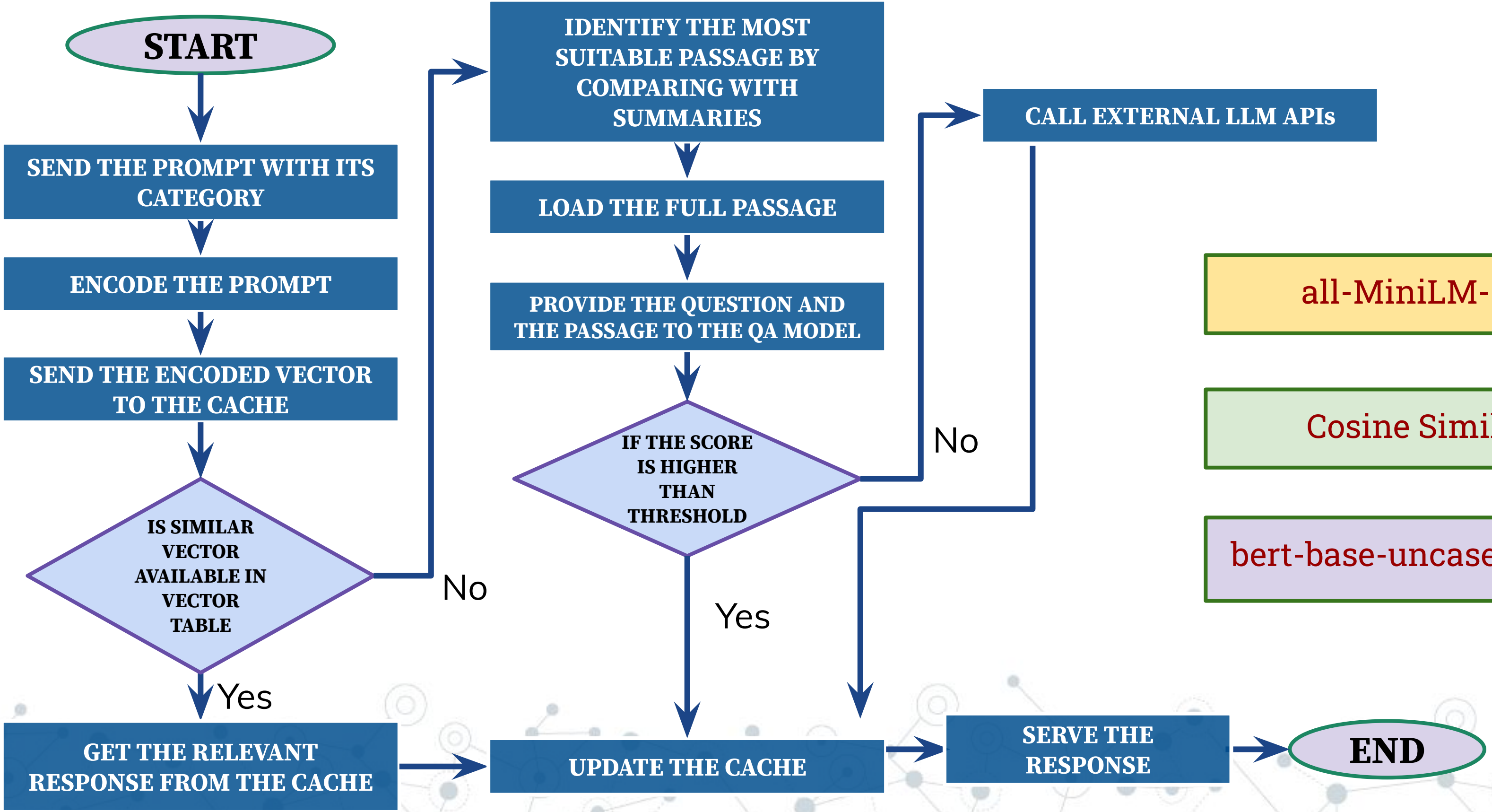
- **A prototype which can be integrated with any module**
  - Cache with a Local Context

3.
Methodology

Collect Course Materials and create datasets

- “Computer architecture” course materials.

Dataflow



START

SEND THE PROMPT WITH ITS CATEGORY

ENCODE THE PROMPT

SEND THE ENCODED VECTOR TO THE CACHE

IS SIMILAR VECTOR AVAILABLE IN VECTOR TABLE

Yes

GET THE RELEVANT RESPONSE FROM THE CACHE

No

IDENTIFY THE MOST SUITABLE PASSAGE BY COMPARING WITH SUMMARIES

LOAD THE FULL PASSAGE

PROVIDE THE QUESTION AND THE PASSAGE TO THE QA MODEL

IF THE SCORE IS HIGHER THAN THRESHOLD

Yes

No

CALL EXTERNAL LLM APIs

UPDATE THE CACHE

SERVE THE RESPONSE

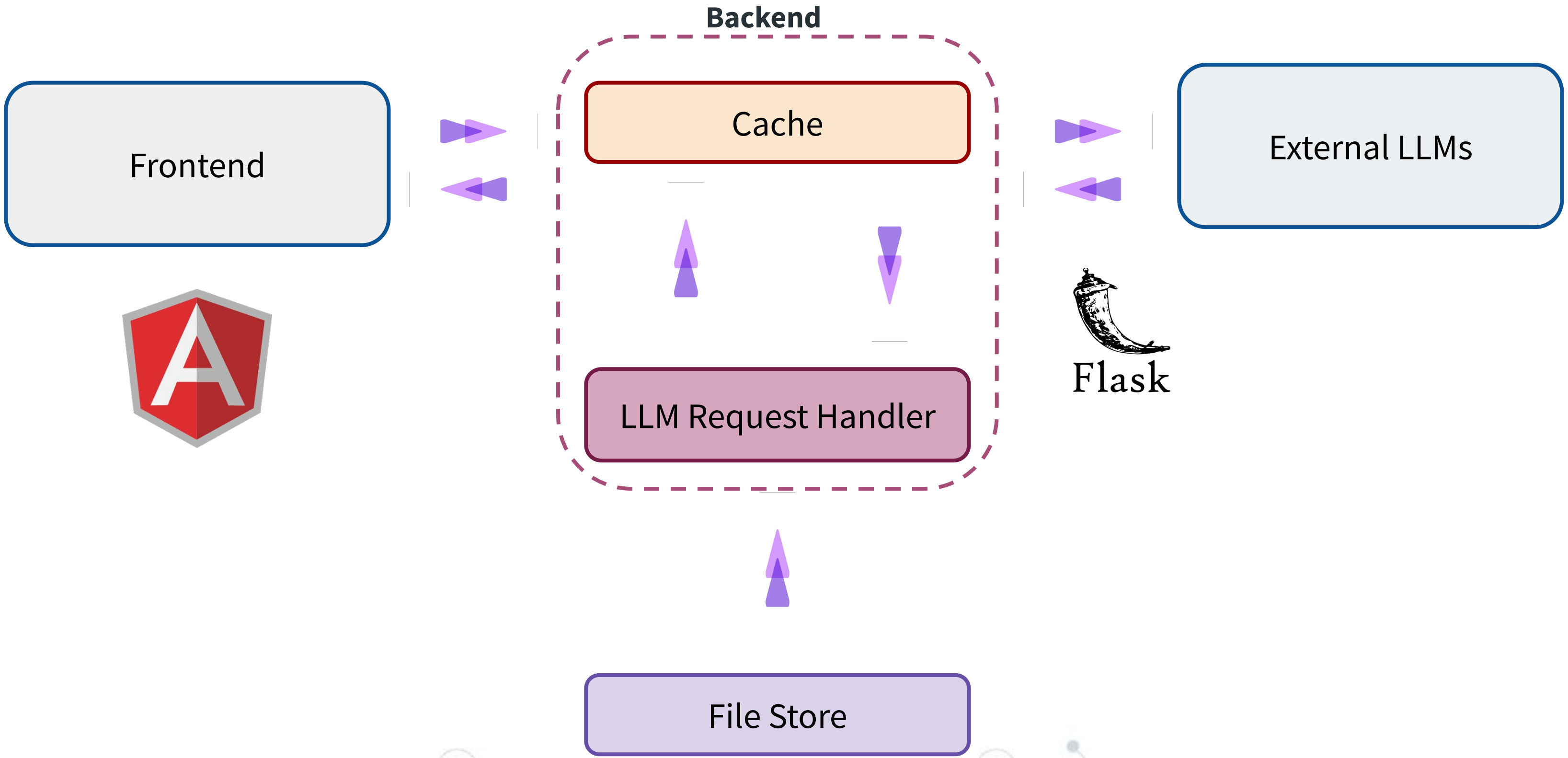
END

all-MiniLM-L6-v2

Cosine Similarity

bert-base-uncased-squad 2

High Level Solution Architecture



Backend

Frontend

Cache

LLM Request Handler

External LLMs

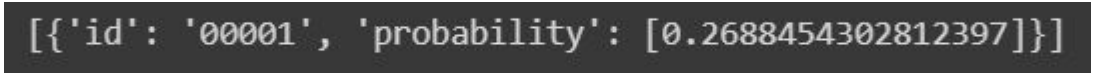
File Store

4.

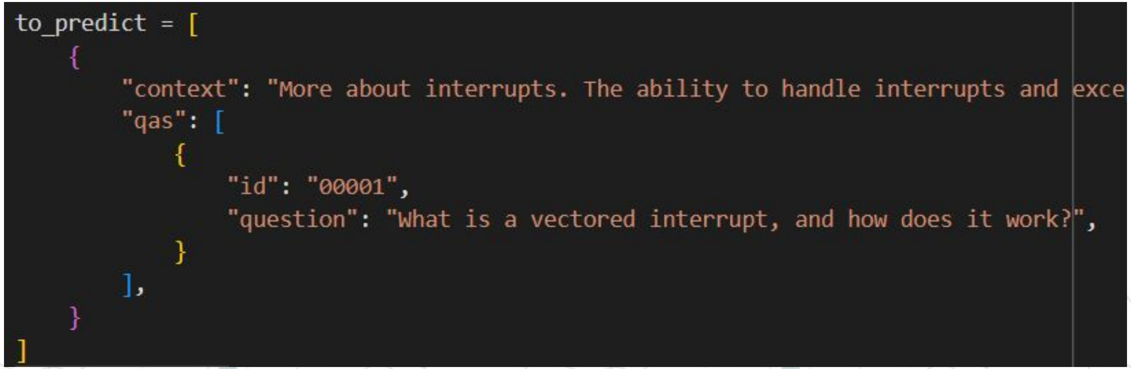
Experiments & Findings

Create Custom models

- bert → bert-base-cased
- electra-base → google/electra-base-discriminator
- roberta → roberta-base
- distilbert → distilbert-base-cased
- distilroberta → distilroberta-base
- electra-small → google/electra-small-discriminator
- xlnet → xlnet-base-cased



```
[{'id': '00001', 'probability': [0.2688454302812397]}]
```



```
to_predict = [
    {
        "context": "More about interrupts. The ability to handle interrupts and exce
        "qas": [
            {
                "id": "00001",
                "question": "What is a vectored interrupt, and how does it work?",
            }
        ],
    }
]
```

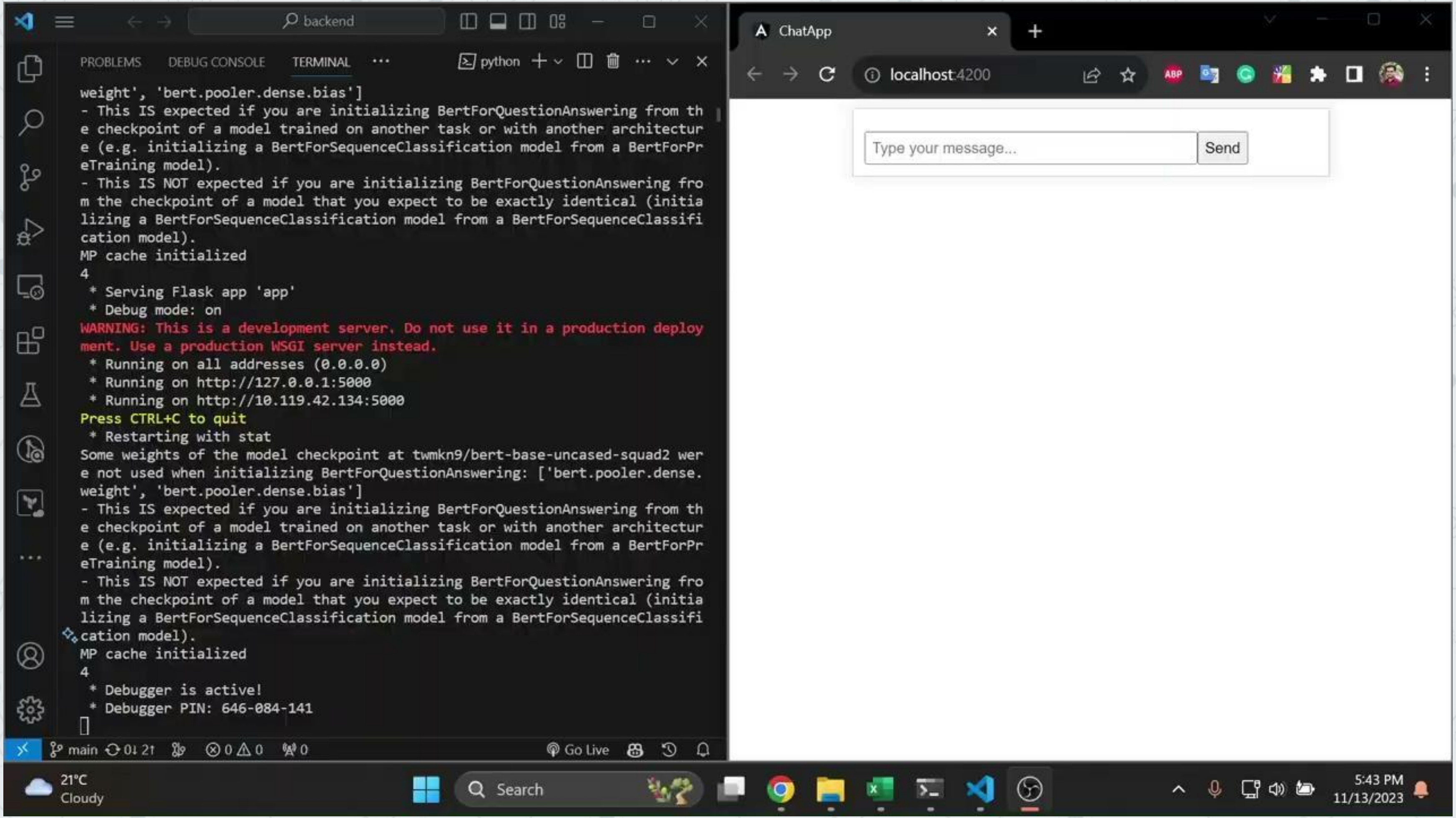
QA Model Implementation

- Use prebuilt Question answering models
- Create custom models to the context

Cache Implementation

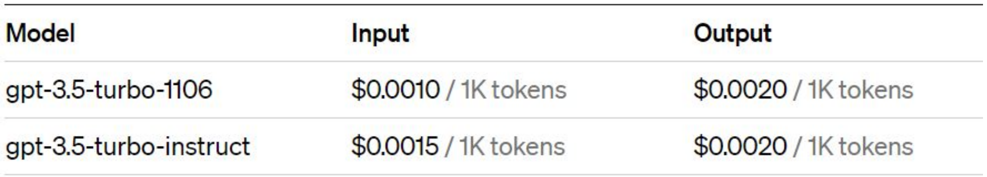
- Test Least Frequency Used eviction policy.
- Test access count when retrieving similar questions.
- Check that cache is updated by the new questions also

3.
Demonstration



```
weight', 'bert.pooler.dense.bias']
- This IS expected if you are initializing BertForQuestionAnswering from th
e checkpoint of a model trained on another task or with another architectur
e (e.g. initializing a BertForSequenceClassification model from a BertForPr
eTraining model).
- This IS NOT expected if you are initializing BertForQuestionAnswering fro
m the checkpoint of a model that you expect to be exactly identical (initia
lizing a BertForSequenceClassification model from a BertForSequenceClassifi
cation model).
MP cache initialized
4
 * Serving Flask app 'app'
 * Debug mode: on
WARNING: This is a development server. Do not use it in a production deploy
ment. Use a production WSGI server instead.
 * Running on all addresses (0.0.0.0)
 * Running on http://127.0.0.1:5000
 * Running on http://10.119.42.134:5000
Press CTRL+C to quit
 * Restarting with stat
Some weights of the model checkpoint at twmkn9/bert-base-uncased-squad2 wer
e not used when initializing BertForQuestionAnswering: ['bert.pooler.dense.
weight', 'bert.pooler.dense.bias']
- This IS expected if you are initializing BertForQuestionAnswering from th
e checkpoint of a model trained on another task or with another architectur
e (e.g. initializing a BertForSequenceClassification model from a BertForPr
eTraining model).
- This IS NOT expected if you are initializing BertForQuestionAnswering fro
m the checkpoint of a model that you expect to be exactly identical (initia
lizing a BertForSequenceClassification model from a BertForSequenceClassifi
cation model).
MP cache initialized
4
 * Debugger is active!
 * Debugger PIN: 646-084-141
```

Cost Analysis



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

Without Our System

No. of Prompts	- 20
No. of API calls	- **20**
Cost per API call	- $0.0030
Total Cost	- **$0.06**

With Our System

No. of Prompts	- 20
No. of API calls	- **6**
Cost per API call	- $0.0030
Total Cost	- **$0.018**

$$\mathit{Cost\ Reduction} = \frac{\$0.06\ -\ \$0.018}{\$0.06}\ \mathrm{x}\ 100 = 70\ \%$$

6.

Deliverables & Their Impact

Deliverables

- A **prototype** which is capable of **integrating with any course materials**

Impact

- Cost reduction
  - Access **Local context** to get the answers
  - **Cache hits** by Similar questions by multiple users

Thank you!

Any Questions?