LLM輔助論文綜整與檢索

組別: 17

112550011李佑軒、112550069謝侑哲、112550183蔡承志

TABLE OF CONTENTS

01

INTRODUCTION

Motivation

找尋適合閱讀的論文非常耗時

GPT4好貴

GPT4只有到2023的資料庫

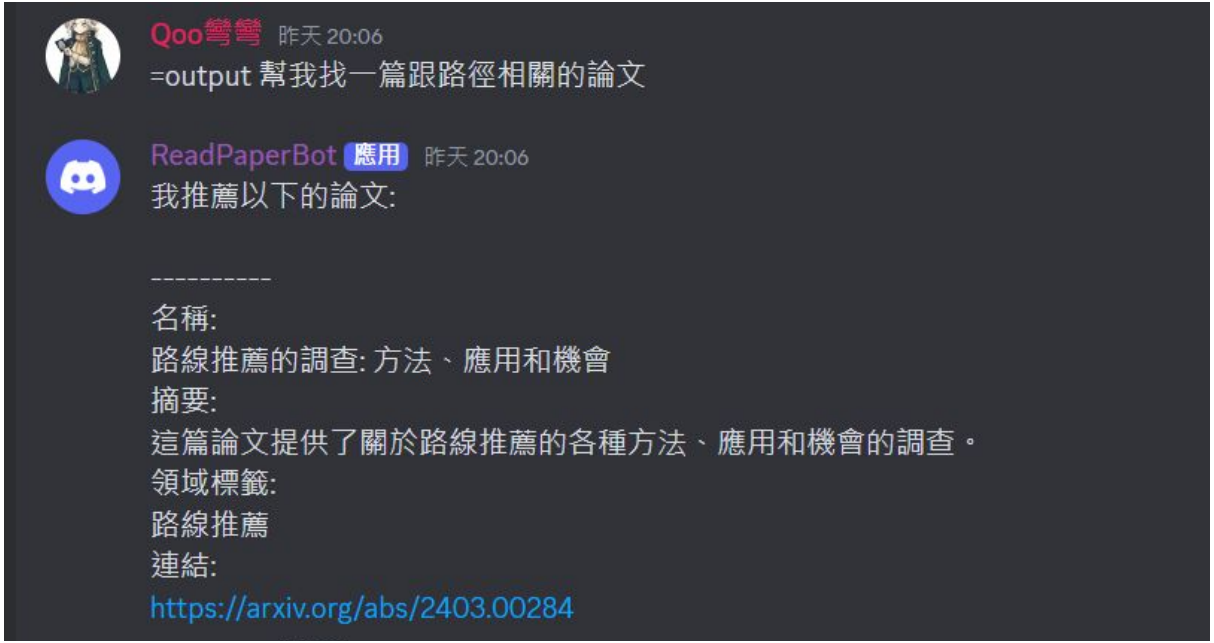
Target

以低成本做出有最新資料庫的LLM-based Discord bot來輔助做論文的檢索以及綜整

Apporach

Crawler: 收集最新論文資料
RAG LLM: 本地資料庫LLM
Discord Bot: 方便互動的介面

User Interface



02
Dataset

選擇使用的論文資料

論文網站：https://arxiv.org/

論文總類：Artificail intelligent

年份範圍：2024 01~06月

總共份數：13107

arXiv > cs.AI

Artificial Intelligence

Notice: Change to 4 digit year in URLs

ArXiv is updating URLs for the /list and /year paths to use 4 digit years: /YYYY for years and /YYYY-MM for months. Old paths will be redirected to the new correct forms wh
year 2002.

Authors and titles for January 2024

Total of 1922 entries : 1-25 26-50 51-75 76-100 ... 1901-1922
Showing up to 25 entries per page: fewer | more | all

[1] arXiv:2401.00004 [pdf, ps, other]
Informational non-reductionist theory of consciousness that providing maximum accuracy of reality prediction
E.E. Vityaev
Comments: 14 pages, 7 figures
Subjects: **Artificial Intelligence (cs.AI)**; Neurons and Cognition (q-bio.NC)

[2] arXiv:2401.00005 [pdf, ps, other]
Consciousness as a logically consistent and prognostic model of reality
Evgenii Vityaev
Comments: 22 pages
Subjects: **Artificial Intelligence (cs.AI)**

[3] arXiv:2401.00006 [pdf, ps, html, other]
Building Open-Ended Embodied Agent via Language-Policy Bidirectional Adaptation
Shaopeng Zhai, Jie Wang, Tianyi Zhang, Fuxian Huang, Qi Zhang, Ming Zhou, Jing Hou, Yu Qiao, Yu Liu
Subjects: **Artificial Intelligence (cs.AI)**

使用爬蟲

arXiv > cs > arXiv:2401.00004

Search...

Help | Adv

Computer Science > Artificial Intelligence

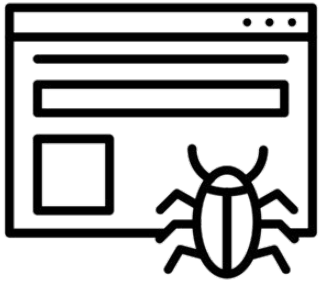
[Submitted on 10 Dec 2023]

Informational non-reductionist theory of consciousness that providing maximum accuracy of reality prediction

E.E. Vityaev

The paper considers a non-reductionist theory of consciousness, which is not reducible to theories of reality and to physiological or psychological theories. Following D.I.Dubrovsky's "informational approach" to the "Mind-Brain Problem", we consider the reality through the prism of information about observed phenomena, which, in turn, is perceived by subjective reality through sensations, perceptions, feelings, etc., which, in turn, are information about the corresponding brain processes. Within this framework the following principle of the Information Theory of Consciousness (ITS) development is put forward: the brain discovers all possible causal relations in the external world and makes all possible inferences by them. The paper shows that ITS built on this principle: (1) also base on the information laws of the structure of external world; (2) explains the structure and functioning of the brain functional systems and cellular ensembles; (3) ensures maximum accuracy of predictions and the anticipation of reality; (4) resolves emerging contradictions and (5) is an information theory of the brain's reflection of reality.

Selenium



存入文字檔

Name
Informational non-reductionist theory of consciousness that providing maximum accuracy of reality prediction

Abstract
The paper considers a non-reductionist theory of consciousness, which is not reducible to theories of reality and to physiological or psychological theories. Following D.I.Dubrovsky's "informational approach" to the "Mind-Brain Problem", we consider the reality through the prism of information about observed phenomena, which, in turn, is perceived by subjective reality through sensations, perceptions, feelings, etc., which, in turn, are information about the corresponding brain processes. Within this framework the following principle of the Information Theory of Consciousness (ITS) development is put forward: the brain discovers all possible causal relations in the external world and makes all possible inferences by them. The paper shows that ITS built on this principle: (1) also base on the information laws of the structure of external world; (2) explains the structure and functioning of the brain functional systems and cellular ensembles; (3) ensures maximum accuracy of predictions and the anticipation of reality; (4) resolves emerging contradictions and (5) is an information theory of the brain's reflection of reality.

Link:
https://arxiv.org/abs/2401.00004

- 13097.txt
- 13098.txt
- 13099.txt
- 13100.txt
- 13101.txt
- 13102.txt
- 13103.txt
- 13104.txt
- 13105.txt
- 13106.txt
- 13107.txt

03

Main Approach - LLM

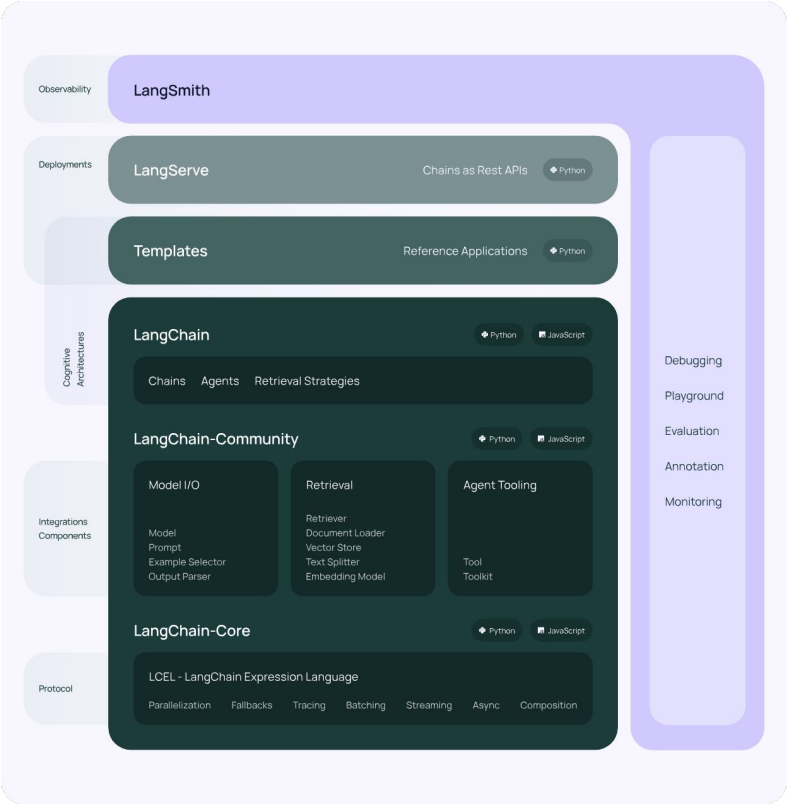
RAG

- 檢索增強生成
- 使用資料庫儲存額外的資料
- 利用transformer-based model
  快速抓出與query相關的資料
- 把抓到的資料連同query餵給LLM

成功只使用pretrained model而不用
tunning, retrain就能使用最新的資料

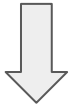
langchain

- framework of LLM
- build a chain contains relevant tools about LLM

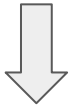


model workflow

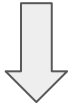
1. 建立RAG用資料庫



2. 建立LLM



3. Prompt Engineering



4. LLM運算及例外處理

```
class Agent:
    def __init__(self):…

    def ask_without_db(self, prompt: str, model="gpt-3.5-turbo"):…

    def ask_question(self, prompt):…
```

Database

langchain tools:

1. document loader
   a. 把dataset的資料load進來
2. text splitter
   a. 把文字分割，使其符合LLM的transformer輸入限制
3. embedding model
   a. embedding文字成向量
4. vector database
   a. Chroma

```
loader = DirectoryLoader('LLM/database', glob='*.txt') # load all the .txt i
documents = loader.load() # type: list{Document}
end = time()
print('time: ', end - start)

print('split text')
start = time()
# text_splitter to split the document to adjust to the input limit of LLM
text_splitter = CharacterTextSplitter(chunk_size=300, chunk_overlap=0)
splited_docs = text_splitter.split_documents(documents) # type: list{Documen
end = time()
print('time: ', end - start)

print('build db')
# embedding text to vector for model
start = time()
embeddings = OpenAIEmbeddings()
# store vector in Chroma vector database
self.db = Chroma.from_documents(splited_docs, embeddings) # type: db
```

LLM

2 kinds of agent:

1. with db: RetrieveQA
2. no db: OpenAI chatmodel

LLM: GPT-3.5-turbo

```
self.qa = RetrievalQA.from_chain_type(llm=ChatOpenAI(model_name='gpt-3.5-turbo'),
                                      chain_type="map_reduce", retriever=self.db.as_retriever(),
                                      return_source_documents=True)
end = time()
print('time: ', end - start)
print('init end')

ask_without_db(self, prompt: str, model="gpt-3.5-turbo"):
print('ask_without_db start')
start = time()
response = openai.OpenAI().chat.completions.create(
    model=model,
    messages=[
        {"role": "user", "content": prompt},
    ]
)
```

Prompt Engineering

5 kinds of prompt engineering:

1. 角色扮演
2. question optimization
3. 中英混合
4. 符號強調
5. few-shoot prompting

```
You are a professional computer science scientist who recommand paper for people
```

```
Please recommand a paper **according to the format of example output**.
Answer it in **English**
用**英文回答**
```

```
----------
Name:
    paper name
Summery:
    summary of paper
Domain Tag:
    tag1, tag2
Link:
    https://arxiv.org/abs/2403.00284
----------
```

outlier handling

LLM with db: 傾向於輸出資料庫查詢結果，否則輸出不知道
If data not found in database -> use LLM without db

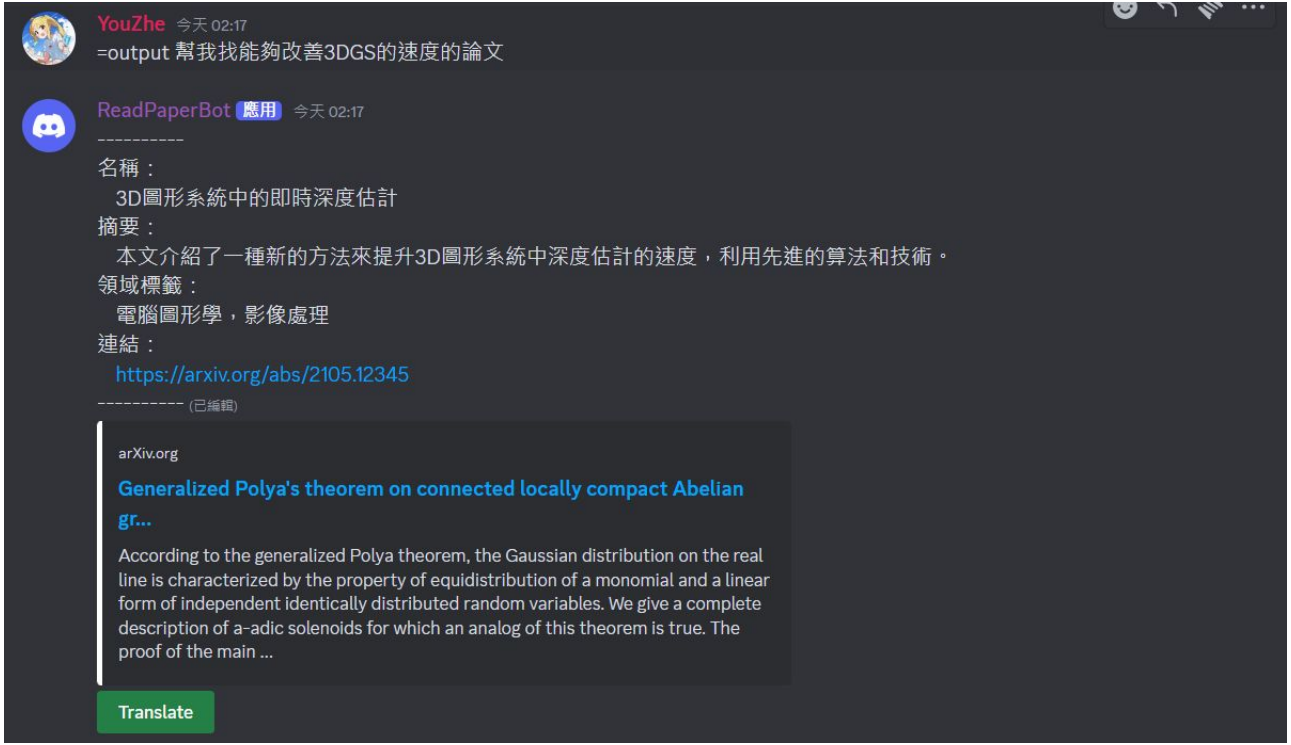
```
if "I don't" in result['result'] or "I do not" in result['result'] or '
    result = self.ask_without_db(query)
else:
    result = result['result']
```

04

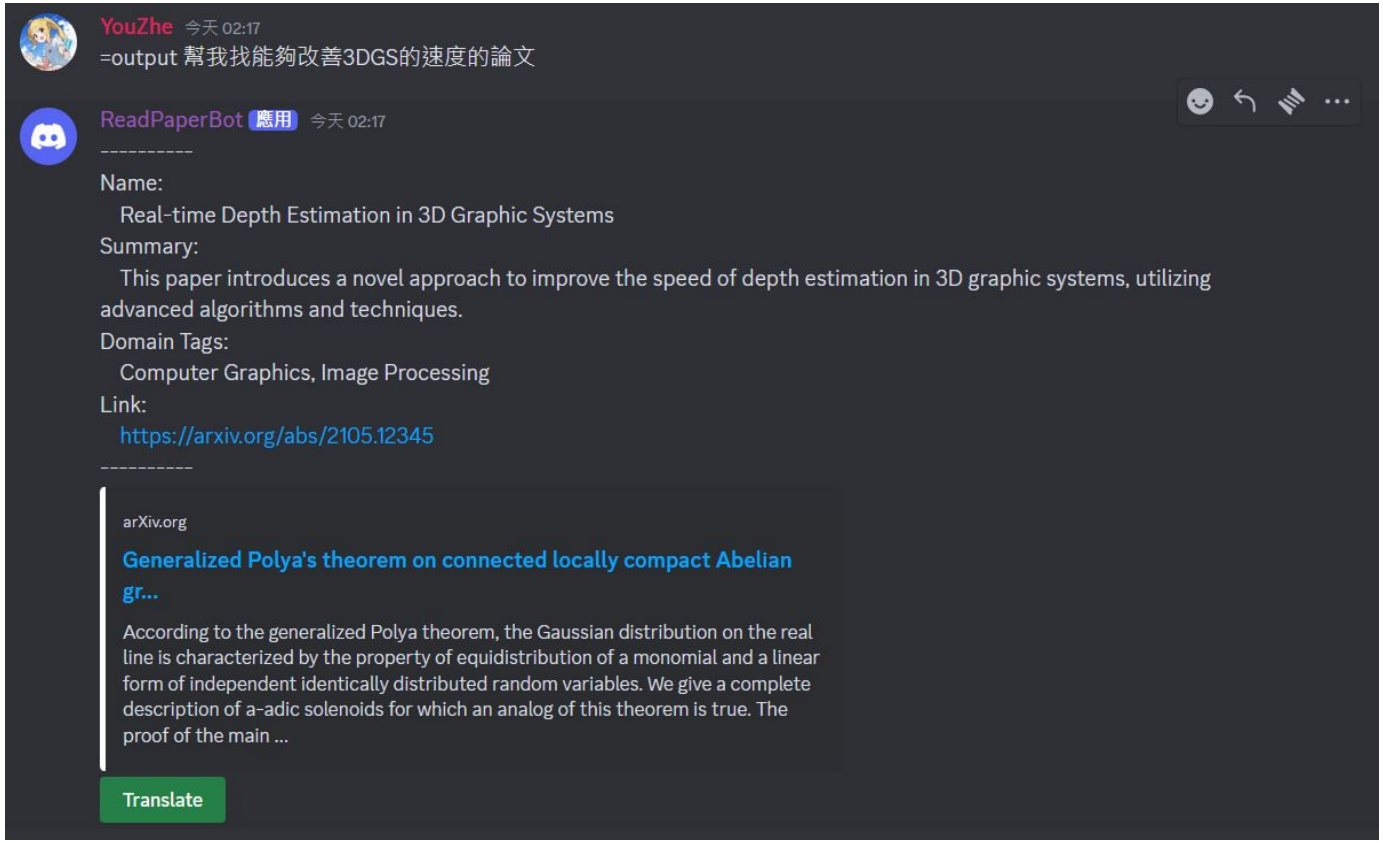
Main Approach - Discord

Discord Bot

- =search 問句
- 輸出格式:
  名稱:
  摘要:
  標籤:
  連結:
- 翻譯Button



Discord Bot (Translate)



05

Result & Analysis

Our Model

1. 格式正確率為93.3%(28/30)
2. 內容正確率為13.3%(4/30)
3. 內容相關率為96.6%(29/30)

Google scholar

1. None
2. None
3. 內容相關率只有6.6%(2/30)

內容正確率為13.3%(4/30) -> 全都是Link出錯

1. Link不能被有意義的解讀 -> 修改資料的儲存方式
2. 能參考的上下文不足 -> 擴增分割的chunk or 取得更精簡的資料

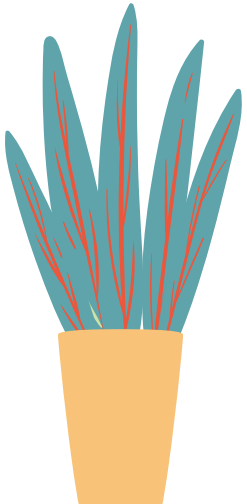
06

Future work

Improve and New function

- Improve:
  - 改善給出不符合的link的問題
  - 輸出格式跑掉
  - 運行時間較久

- New function:
  - 論文的pdf檔做完整解析與總結
  - 自動擴充資料庫
  - 介面優化



Thanks