Easy 100% Local RAG Tutorial (Ollama) + Full Code

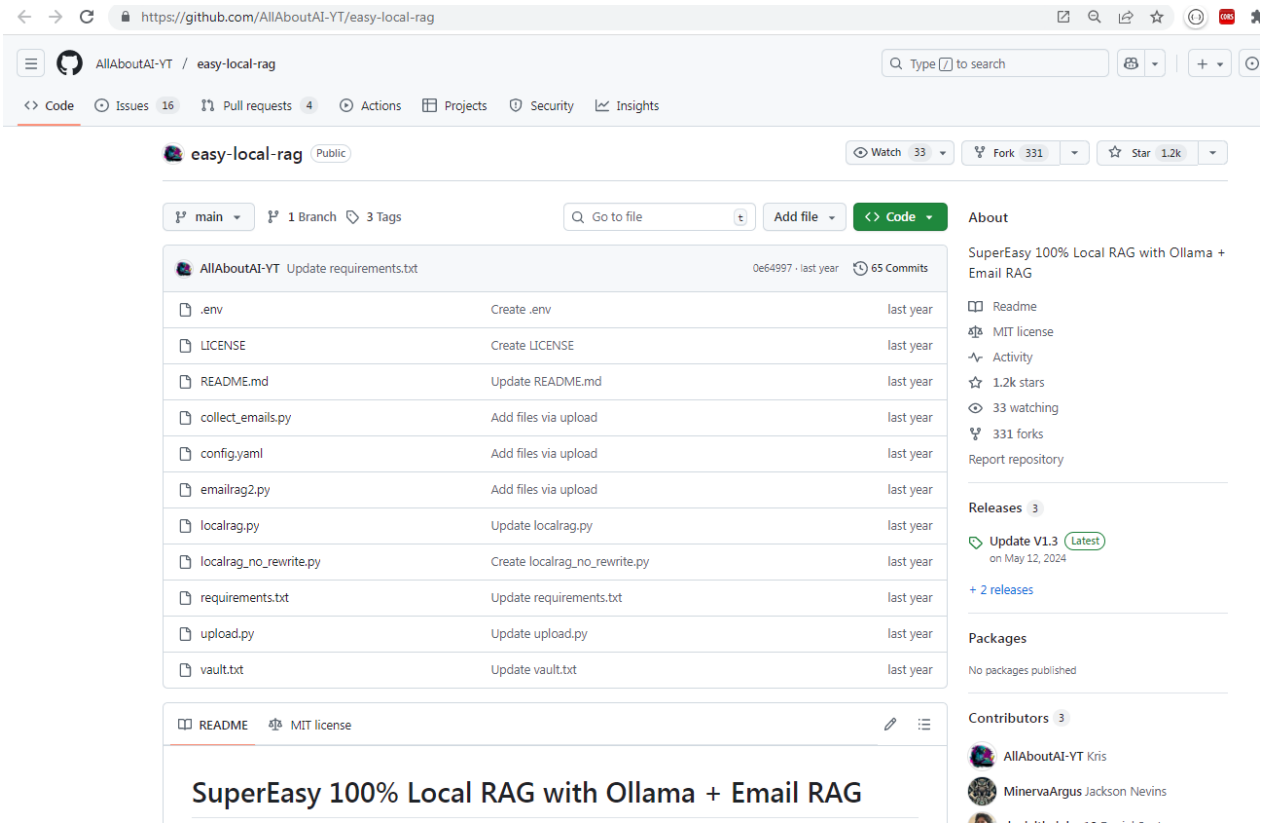
All About AI
215K subscribers

Join Subscribe 2.1K

96,464 views Apr 15, 2024
Easy 100% Local RAG Tutorial (Ollama) + Full Code

In this video I create a 100% local RAG in around 70 lines of code. Feel free to share and rate on GitHub :)

00:00 Local RAG Intro
02:01 Local RAG Full Tutorial



Setup

1. git clone https://github.com/AllAboutAI-YT/easy-local-rag.git
2. cd dir
3. pip install -r requirements.txt
4. Install Ollama (https://ollama.com/download)
5. ollama pull llama3 (etc)
6. ollama pull mxbai-embed-large
7. run upload.py (pdf, .txt, JSON)
8. run localrag.py (with query re-write)
9. run localrag_no_rewrite.py (no query re-write)

Email RAG Setup

1. git clone https://github.com/AllAboutAI-YT/easy-local-rag.git
2. cd dir
3. pip install -r requirements.txt
4. Install Ollama (https://ollama.com/download)
5. ollama pull llama3 (etc)
6. ollama pull mxbai-embed-large
7. set YOUR email logins in .env (for gmail create app password (video))
8. python collect_emails.py to download your emails
9. python emailrag2.py to talk to your emails

Latest Updates

- Added Email RAG Support (v1.3)
- Upload.py (v1.2)
  - replaced /n/n with /n
- New embeddings model mxbai-embed-large from ollama (1.2)
- Rewrite query function to improve retrival on vauge questions (1.2)
- Pick your model from the CLI (1.1)
  - python localrag.py --model mistral (llama3 is default)
- Talk in a true loop with conversation history (1.1)

My YouTube Channel

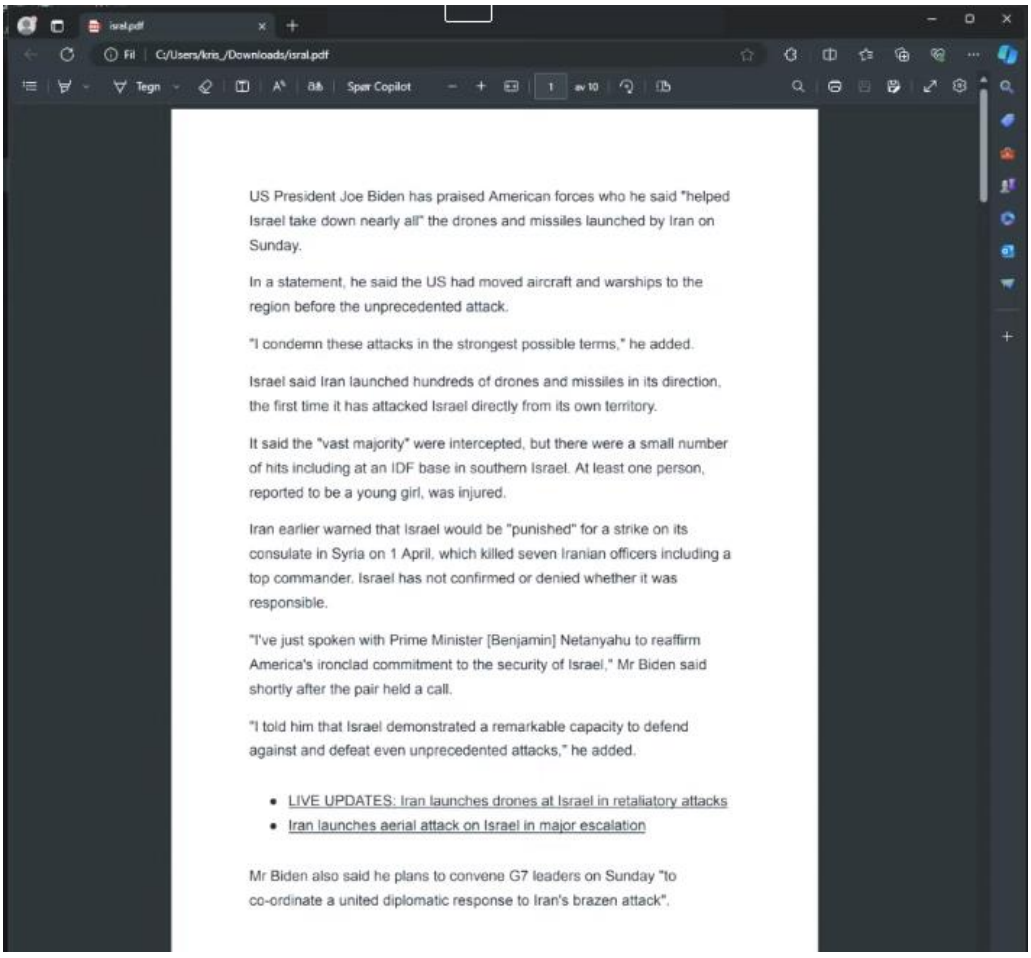
https://www.youtube.com/c/AllAboutAI

What is RAG?

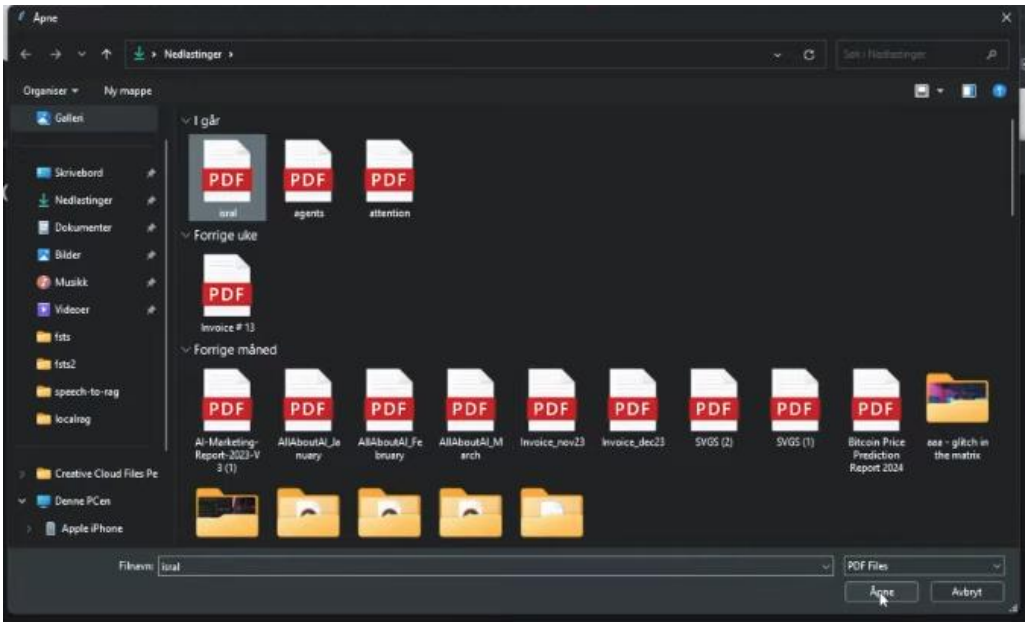
RAG is a way to enhance the capabilities of LLMs by combining their powerful language understanding with targeted retrieval of relevant information from external sources often with using embeddings in vector databases, leading to more accurate, trustworthy, and versatile AI-powered applications

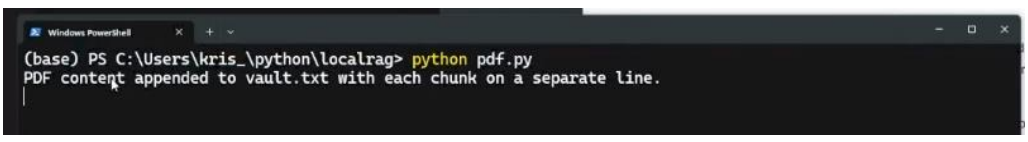
What is Ollama?

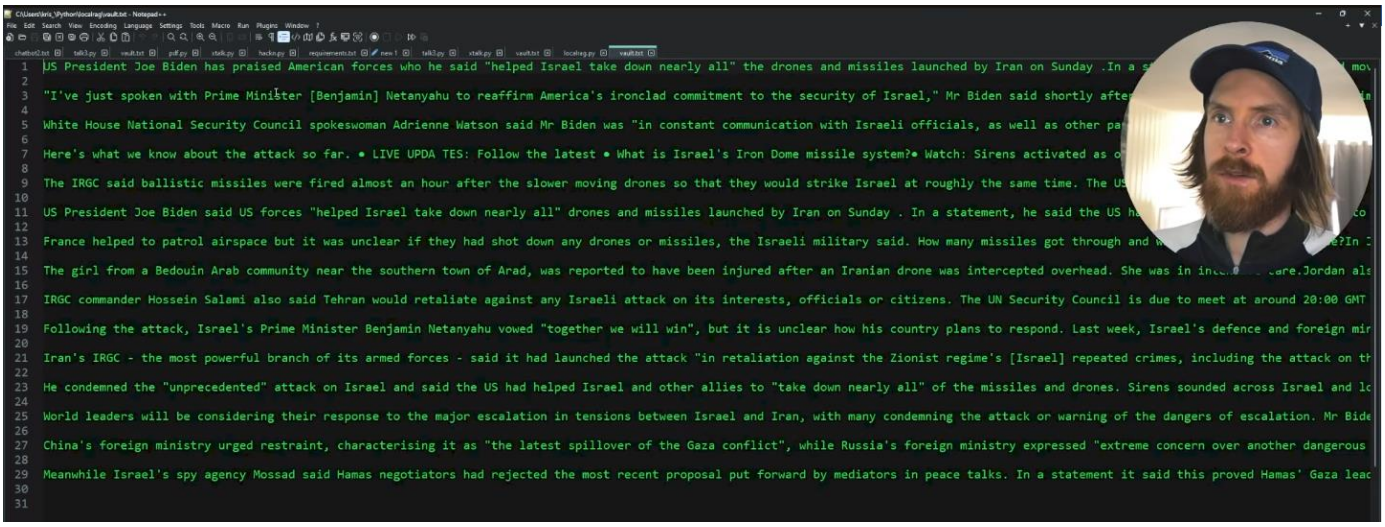
Ollama is an open-source platform that simplifies the process of running powerful LLMs locally on your own machine, giving users more control and flexibility in their AI projects. https://www.ollama.com


isral.pdf
C:/Users/kris_/Downloads/isral.pdf
Tegn
Spør Copilot
US President Joe Biden has praised American forces who he said "helped Israel take down nearly all" the drones and missiles launched by Iran on Sunday.
In a statement, he said the US had moved aircraft and warships to the region before the unprecedented attack.
"I condemn these attacks in the strongest possible terms," he added.
Israel said Iran launched hundreds of drones and missiles in its direction, the first time it has attacked Israel directly from its own territory.
It said the "vast majority" were intercepted, but there were a small number of hits including at an IDF base in southern Israel. At least one person, reported to be a young girl, was injured.
Iran earlier warned that Israel would be "punished" for a strike on its consulate in Syria on 1 April, which killed seven Iranian officers including a top commander. Israel has not confirmed or denied whether it was responsible.
"I've just spoken with Prime Minister [Benjamin] Netanyahu to reaffirm America's ironclad commitment to the security of Israel," Mr Biden said shortly after the pair held a call.
"I told him that Israel demonstrated a remarkable capacity to defend against and defeat even unprecedented attacks," he added.
LIVE UPDATES: Iran launches drones at Israel in retaliatory attacks
Iran launches aerial attack on Israel in major escalation
Mr Biden also said he plans to convene G7 leaders on Sunday "to co-ordinate a united diplomatic response to Iran's brazen attack".

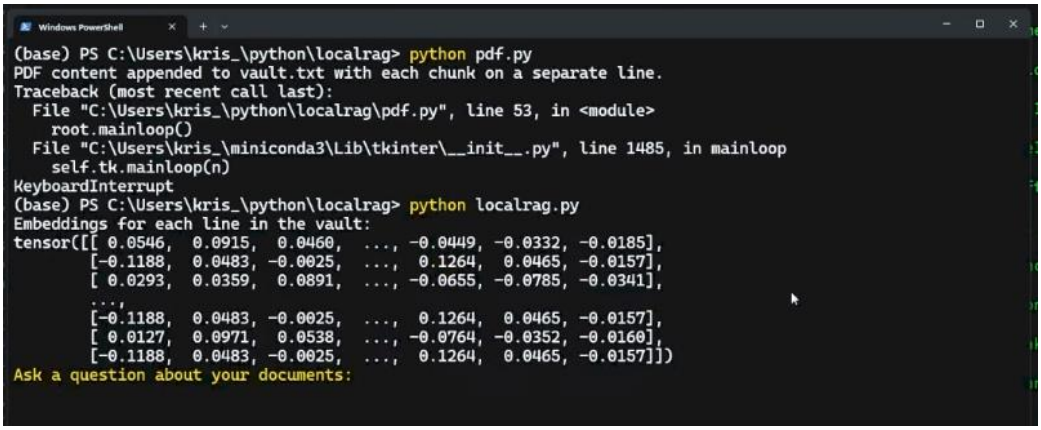

Windows PowerShell
(base) PS C:\Users\kris_\python\localrag> python pdf.py


Åpne
Nedlastinger
Organiser
Ny mappe
I går
isral
agents
attention
Forrige uke
Invoice # 13
Forrige måned
AI-Marketing-Report-2023-V3 (1)
AllAboutAI_January
AllAboutAI_February
AllAboutAI_March
Invoice_nov23
Invoice_dec23
SVGS (2)
SVGS (1)
Bitcoin Price Prediction Report 2024
Filnavn: isral
PDF Files
Åpne
Avbryt

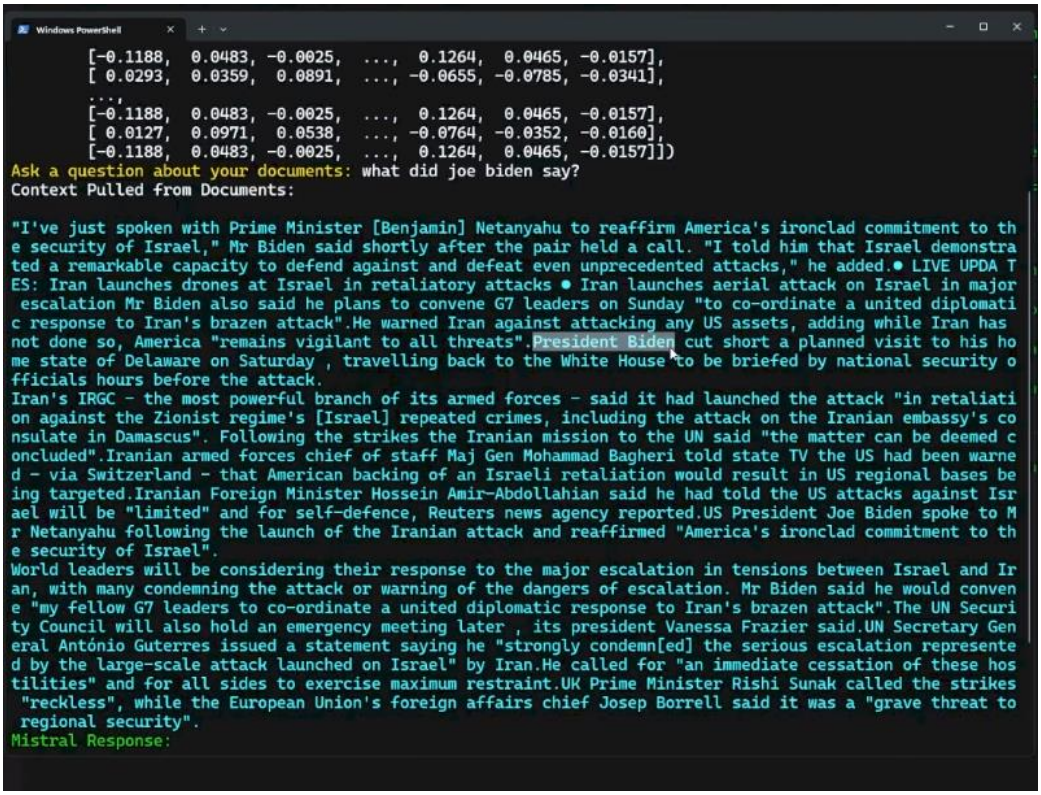

Windows PowerShell
(base) PS C:\Users\kris_\python\localrag> python pdf.py
PDF content appended to vault.txt with each chunk on a separate line.



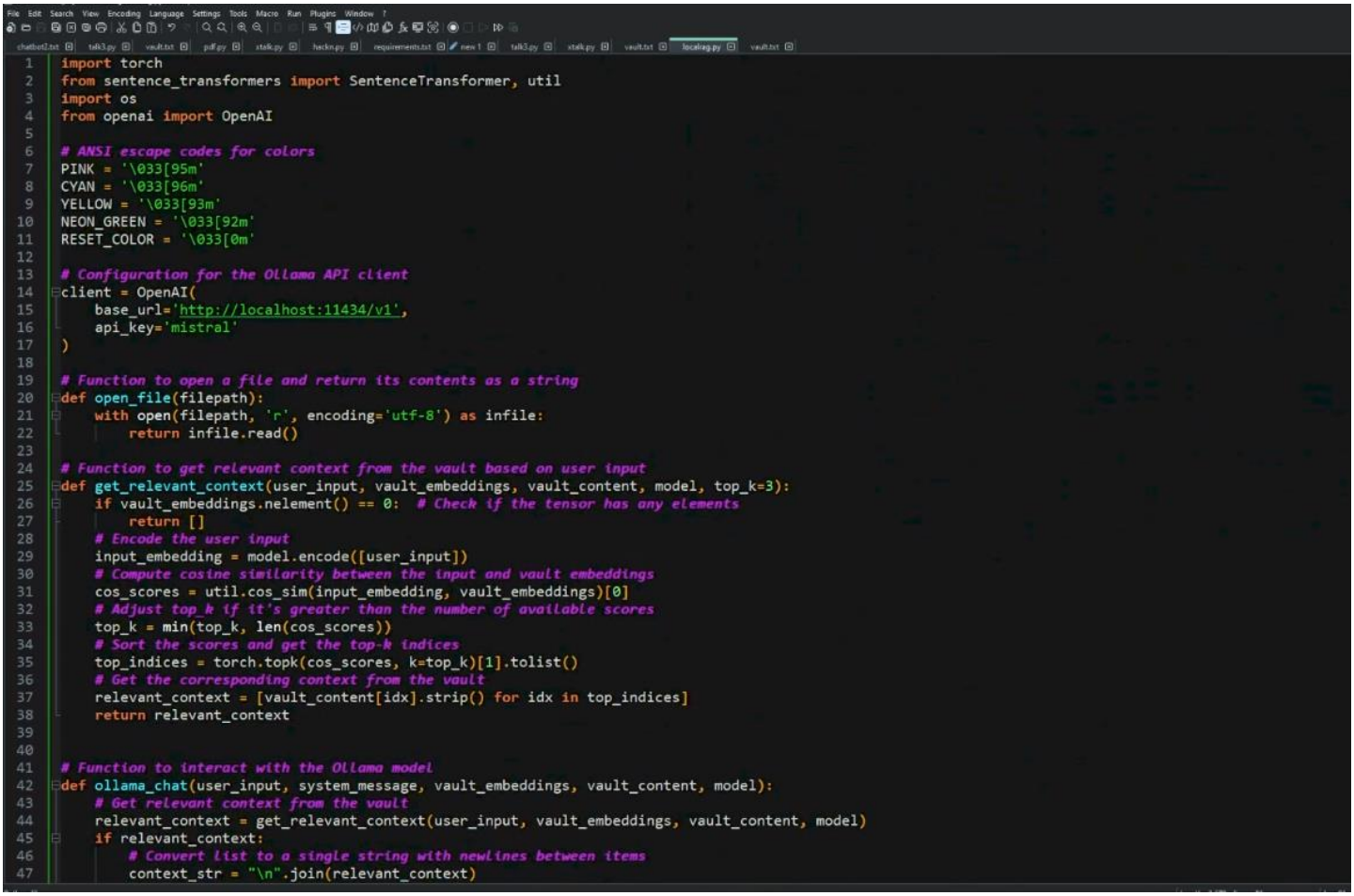
We want our chunks on separate lines as above



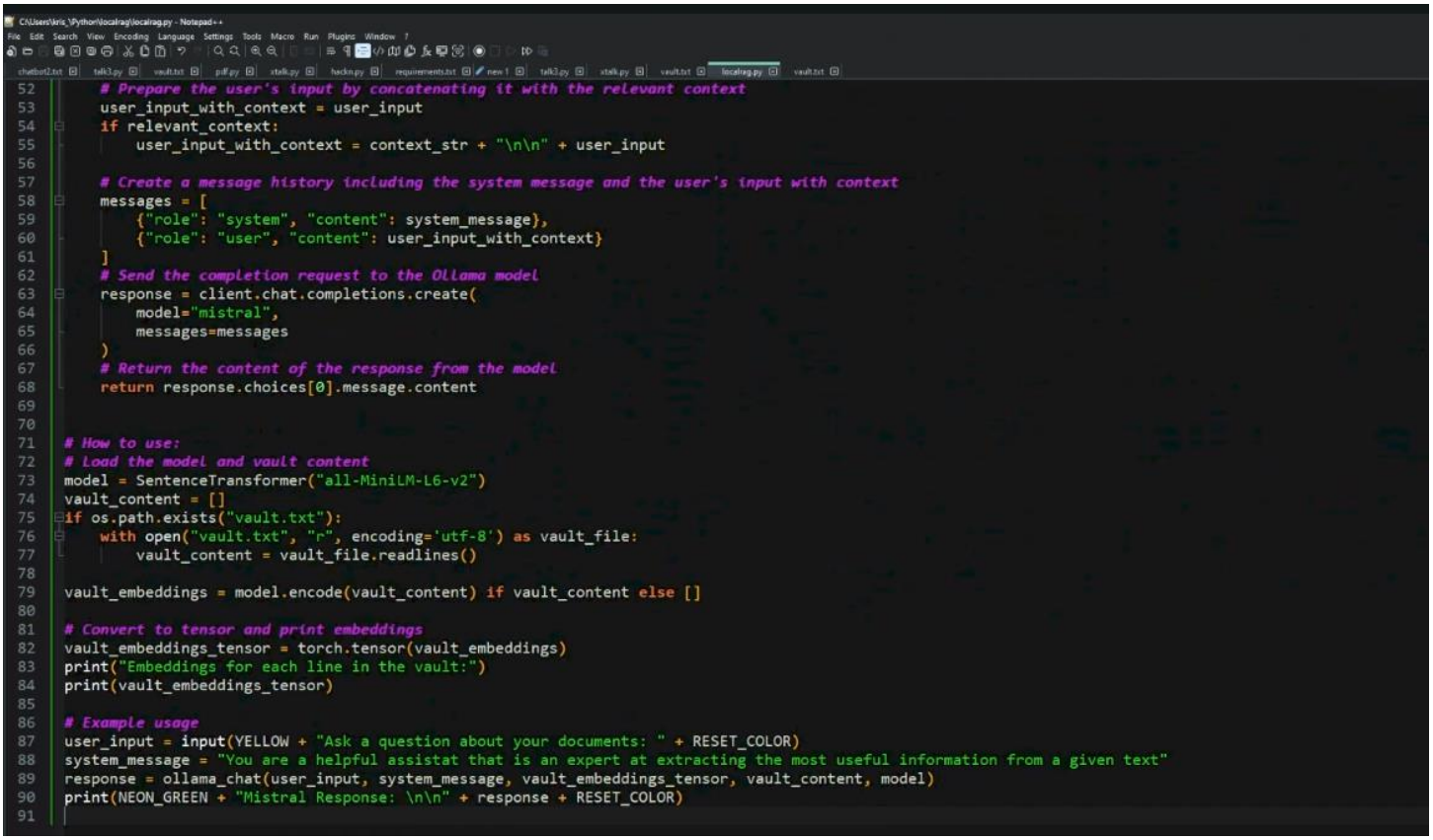
We create the chunk embeddings and can now start asking questions about our document




Windows PowerShell
"I've just spoken with Prime Minister [Benjamin] Netanyahu to reaffirm America's ironclad commitment to th
e security of Israel," Mr Biden said shortly after the pair held a call. "I told him that Israel demonstra
ted a remarkable capacity to defend against and defeat even unprecedented attacks," he added.● LIVE UPDA T
ES: Iran launches drones at Israel in retaliatory attacks ● Iran launches aerial attack on Israel in major
escalation Mr Biden also said he plans to convene G7 leaders on Sunday "to co-ordinate a united diplomati
c response to Iran's brazen attack".He warned Iran against attacking any US assets, adding while Iran has
not done so, America "remains vigilant to all threats".President Biden cut short a planned visit to his ho
me state of Delaware on Saturday , travelling back to the White House to be briefed by national security o
fficials hours before the attack.
Iran's IRGC - the most powerful branch of its armed forces - said it had launched the attack "in retaliati
on against the Zionist regime's [Israel] repeated crimes, including the attack on the Iranian embassy's co
nsulate in Damascus". Following the strikes the Iranian mission to the UN said "the matter can be deemed c
oncluded".Iranian armed forces chief of staff Maj Gen Mohammad Bagheri told state TV the US had been warne
d - via Switzerland - that American backing of an Israeli retaliation would result in US regional bases be
ing targeted.Iranian Foreign Minister Hossein Amir-Abdollahian said he had told the US attacks against Isr
ael will be "limited" and for self-defence, Reuters news agency reported.US President Joe Biden spoke to M
r Netanyahu following the launch of the Iranian attack and reaffirmed "America's ironclad commitment to th
e security of Israel".
World leaders will be considering their response to the major escalation in tensions between Israel and Ir
an, with many condemning the attack or warning of the dangers of escalation. Mr Biden said he would conven
e "my fellow G7 leaders to co-ordinate a united diplomatic response to Iran's brazen attack".The UN Securi
ty Council will also hold an emergency meeting later , its president Vanessa Frazier said.UN Secretary Gen
eral António Guterres issued a statement saying he "strongly condemn[ed] the serious escalation represente
d by the large-scale attack launched on Israel" by Iran.He called for "an immediate cessation of these hos
tilities" and for all sides to exercise maximum restraint.UK Prime Minister Rishi Sunak called the strikes
"reckless", while the European Union's foreign affairs chief Josep Borrell said it was a "grave threat to
regional security".
Mistral Response:
Joe Biden spoke with Israeli Prime Minister Benjamin Netanyahu to reaffirm America's commitment to Israel
's security following Iran's attack. He also planned to convene G7 leaders to coordinate a united diplomat
ic response to Iran's actions. Biden warned Iran against attacking any US assets and cut short his visit t
o Delaware to be briefed by national security officials. The UN Security Council and many world leaders, i
ncluding the UK Prime Minister Rishi Sunak and the European Union's foreign affairs chief Josep Borrell, c
ondemned the attack and warned of the dangers of escalation.
(base) PS C:\Users\kris_\python\localrag>

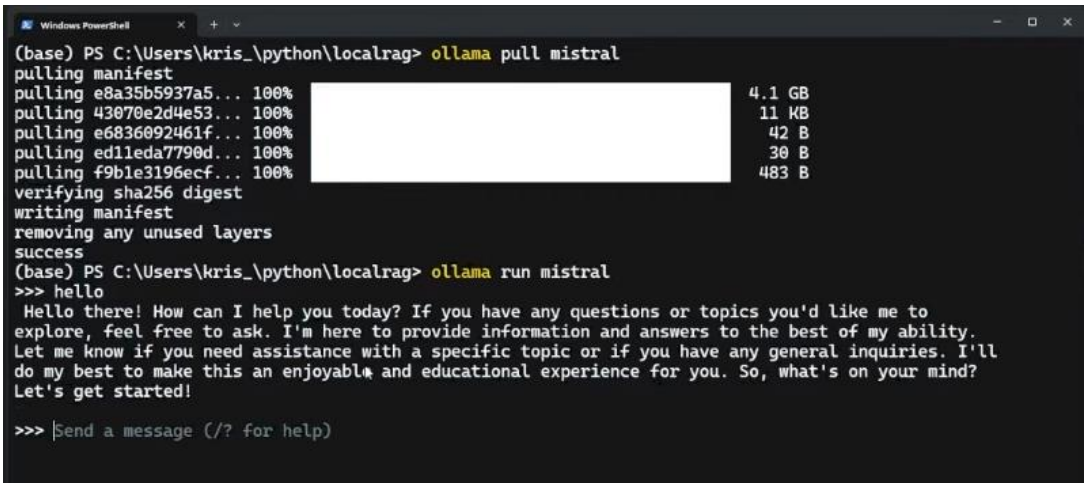

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
chatbot2.txt talk3.py vault.txt pdf.py talk.py hackn.py requirements.txt new 1 talk3.py talk.py vault.txt localrag.py vault.txt
1 import torch
2 from sentence_transformers import SentenceTransformer, util
3 import os
4 from openai import OpenAI
5
6 # ANSI escape codes for colors
7 PINK = '\033[95m'
8 CYAN = '\033[96m'
9 YELLOW = '\033[93m'
10 NEON_GREEN = '\033[92m'
11 RESET_COLOR = '\033[0m'
12
13 # Configuration for the Ollama API client
14 client = OpenAI(
15 base_url='http://localhost:11434/v1',
16 api_key='mistral'
17)
18
19 # Function to open a file and return its contents as a string
20 def open_file(filepath):
21 with open(filepath, 'r', encoding='utf-8') as infile:
22 return infile.read()
23
24 # Function to get relevant context from the vault based on user input
25 def get_relevant_context(user_input, vault_embeddings, vault_content, model, top_k=3):
26 if vault_embeddings.nelement() == 0: # Check if the tensor has any elements
27 return []
28 # Encode the user input
29 input_embedding = model.encode([user_input])
30 # Compute cosine similarity between the input and vault embeddings
31 cos_scores = util.cos_sim(input_embedding, vault_embeddings)[0]
32 # Adjust top_k if it's greater than the number of available scores
33 top_k = min(top_k, len(cos_scores))
34 # Sort the scores and get the top-k indices
35 top_indices = torch.topk(cos_scores, k=top_k)[1].tolist()
36 # Get the corresponding context from the vault
37 relevant_context = [vault_content[idx].strip() for idx in top_indices]
38 return relevant_context
39
40
41 # Function to interact with the Ollama model
42 def ollama_chat(user_input, system_message, vault_embeddings, vault_content, model):
43 # Get relevant context from the vault
44 relevant_context = get_relevant_context(user_input, vault_embeddings, vault_content, model)
45 if relevant_context:
46 # Convert list to a single string with newlines between items
47 context_str = "\n".join(relevant_context)

```python
# Function to interact with the Ollama model
def ollama_chat(user_input, system_message, vault_embeddings, vault_content, model):
    # Get relevant context from the vault
    relevant_context = get_relevant_context(user_input, vault_embeddings, vault_content, model)
    if relevant_context:
        # Convert list to a single string with newlines between items
        context_str = "\n".join(relevant_context)
        print("Context Pulled from Documents: \n\n" + CYAN + context_str + RESET_COLOR)
    else:
        print(CYAN + "No relevant context found." + RESET_COLOR)

    # Prepare the user's input by concatenating it with the relevant context
    user_input_with_context = user_input
    if relevant_context:
        user_input_with_context = context_str + "\n\n" + user_input

    # Create a message history including the system message and the user's input with context
    messages = [
        {"role": "system", "content": system_message},
        {"role": "user", "content": user_input_with_context}
    ]
    # Send the completion request to the Ollama model
    response = client.chat.completions.create(
        model="mistral",
```

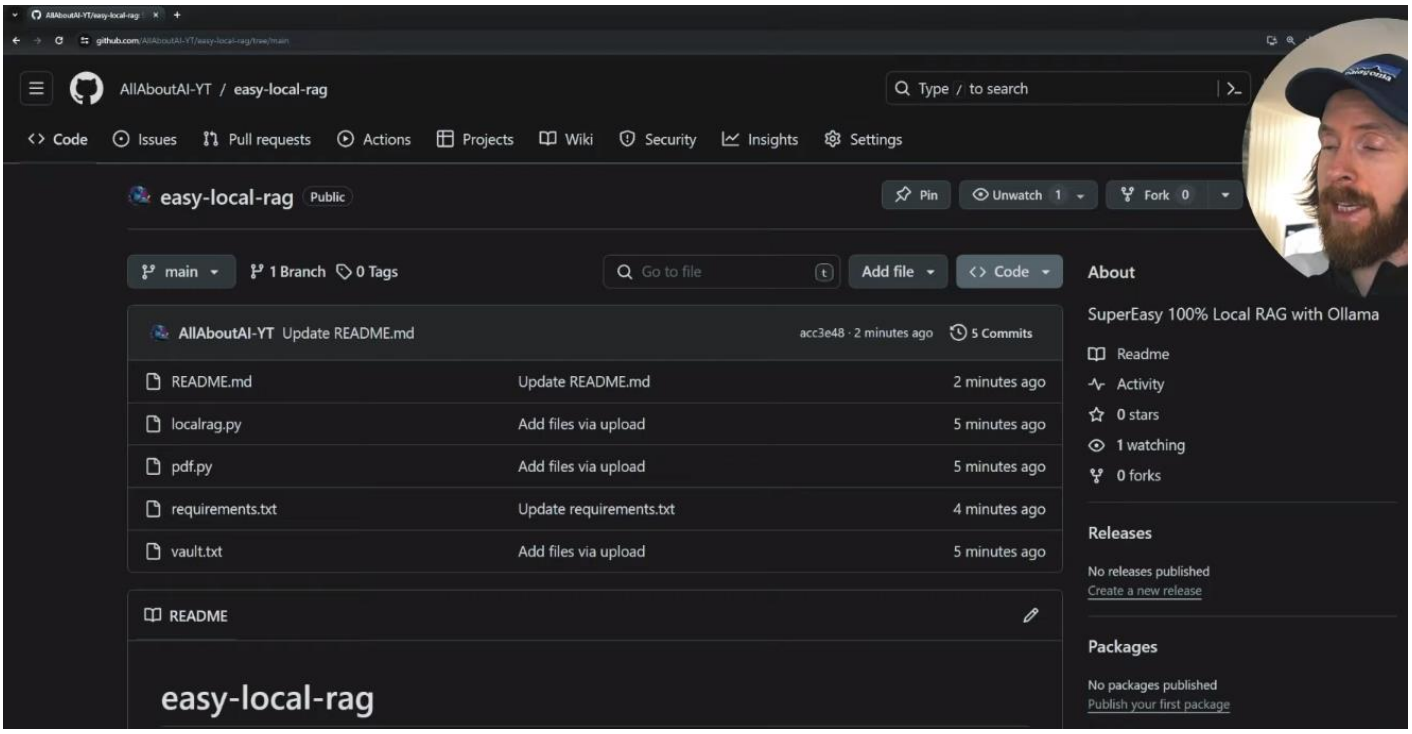


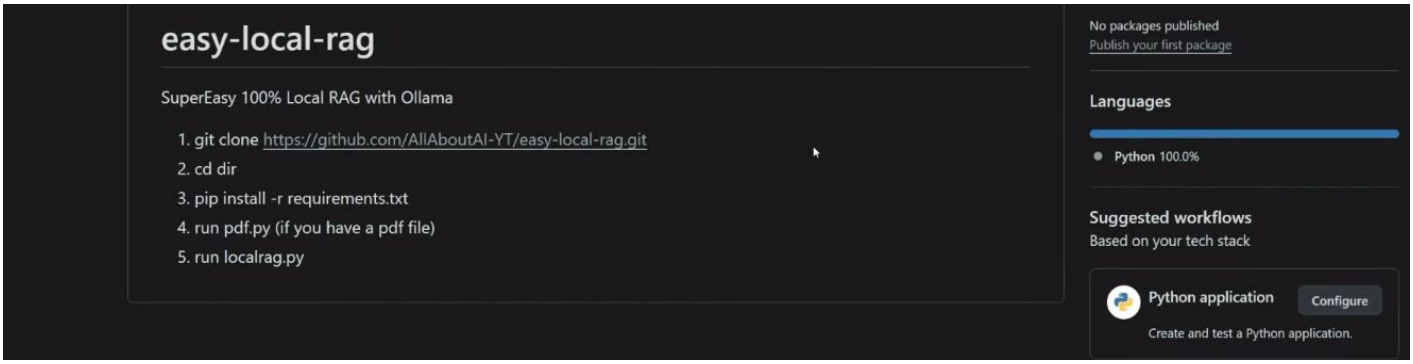
```python
    # Prepare the user's input by concatenating it with the relevant context
    user_input_with_context = user_input
    if relevant_context:
        user_input_with_context = context_str + "\n\n" + user_input

    # Create a message history including the system message and the user's input with context
    messages = [
        {"role": "system", "content": system_message},
        {"role": "user", "content": user_input_with_context}
    ]
    # Send the completion request to the Ollama model
    response = client.chat.completions.create(
        model="mistral",
        messages=messages
    )
    # Return the content of the response from the model
    return response.choices[0].message.content


# How to use:
# Load the model and vault content
model = SentenceTransformer("all-MiniLM-L6-v2")
vault_content = []
if os.path.exists("vault.txt"):
    with open("vault.txt", "r", encoding='utf-8') as vault_file:
        vault_content = vault_file.readlines()

vault_embeddings = model.encode(vault_content) if vault_content else []

# Convert to tensor and print embeddings
vault_embeddings_tensor = torch.tensor(vault_embeddings)
print("Embeddings for each line in the vault:")
print(vault_embeddings_tensor)

# Example usage
user_input = input(YELLOW + "Ask a question about your documents: " + RESET_COLOR)
system_message = "You are a helpful assistat that is an expert at extracting the most useful information from a given text"
response = ollama_chat(user_input, system_message, vault_embeddings_tensor, vault_content, model)
print(NEON_GREEN + "Mistral Response: \n\n" + response + RESET_COLOR)
```

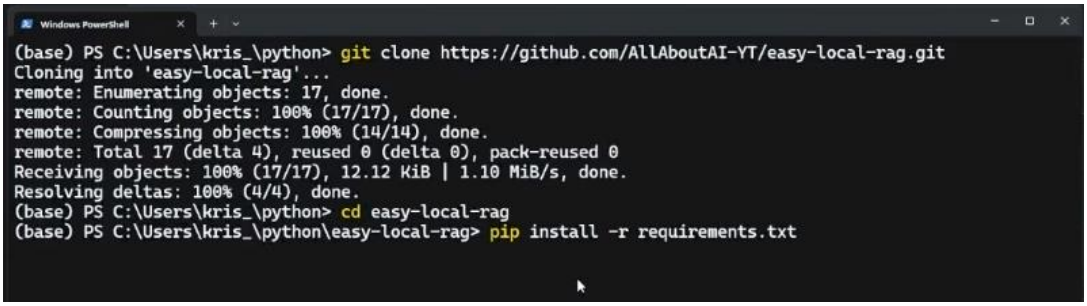
This is the main code.

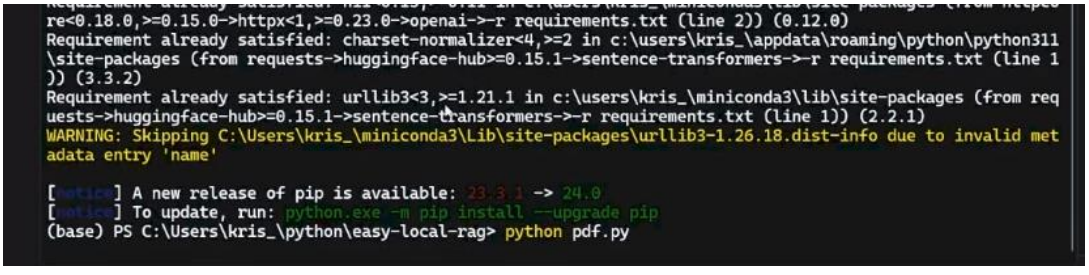


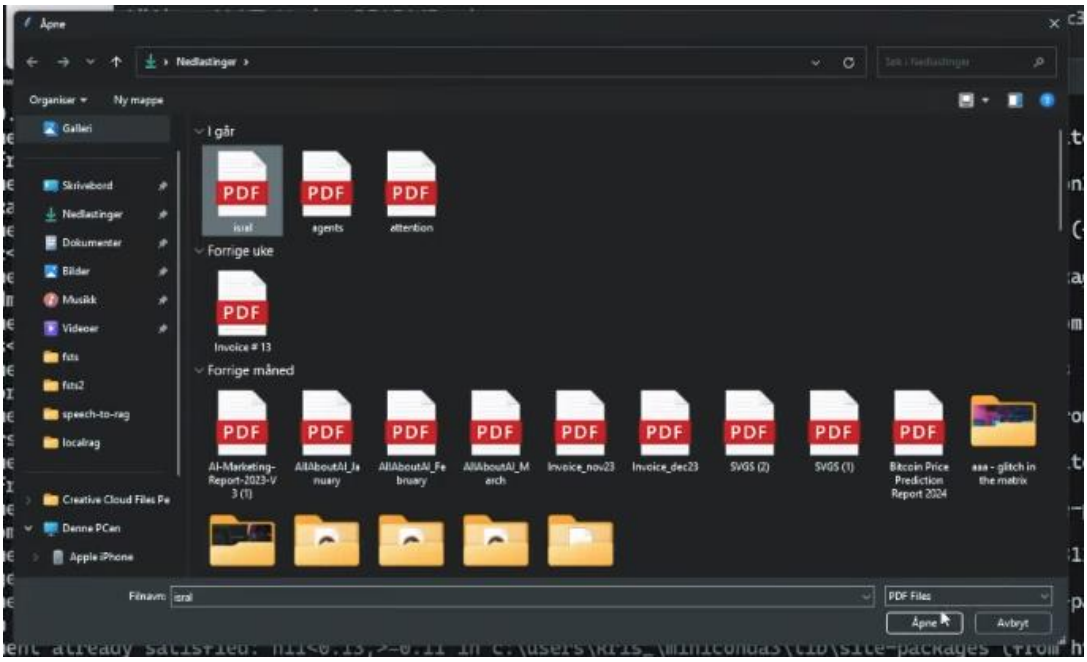
```
Windows PowerShell
(base) PS C:\Users\kris_\python\localrag> ollama pull mistral
pulling manifest
pulling e8a35b5937a5... 100%  4.1 GB
pulling 43070e2d4e53... 100%   11 KB
pulling e6836092461f... 100%   42 B
pulling ed11eda7790d... 100%   30 B
pulling f9b1e3196ecf... 100%   483 B
verifying sha256 digest
writing manifest
removing any unused layers
success
(base) PS C:\Users\kris_\python\localrag> ollama run mistral
>>> hello
 Hello there! How can I help you today? If you have any questions or topics you'd like me to
explore, feel free to ask. I'm here to provide information and answers to the best of my ability.
Let me know if you need assistance with a specific topic or if you have any general inquiries. I'll
do my best to make this an enjoyable and educational experience for you. So, what's on your mind?
Let's get started!

>>> Send a message (/? for help)
```


AllAboutAI-YT / easy-local-rag
Type / to search
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
easy-local-rag
Public
Pin
Unwatch 1
Fork 0
main
1 Branch
0 Tags
Go to file
Add file
Code
AllAboutAI-YT Update README.md
acc3e48 · 2 minutes ago
5 Commits
README.md Update README.md 2 minutes ago
localrag.py Add files via upload 5 minutes ago
pdf.py Add files via upload 5 minutes ago
requirements.txt Update requirements.txt 4 minutes ago
vault.txt Add files via upload 5 minutes ago
README
easy-local-rag
About
SuperEasy 100% Local RAG with Ollama
Readme
Activity
0 stars
1 watching
0 forks
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package


easy-local-rag
SuperEasy 100% Local RAG with Ollama
1. git clone https://github.com/AllAboutAI-YT/easy-local-rag.git
2. cd dir
3. pip install -r requirements.txt
4. run pdf.py (if you have a pdf file)
5. run localrag.py
No packages published
Publish your first package
Languages
Python 100.0%
Suggested workflows
Based on your tech stack
Python application
Configure
Create and test a Python application.


Windows PowerShell
(base) PS C:\Users\kris_\python> git clone https://github.com/AllAboutAI-YT/easy-local-rag.git
Cloning into 'easy-local-rag'...
remote: Enumerating objects: 17, done.
remote: Counting objects: 100% (17/17), done.
remote: Compressing objects: 100% (14/14), done.
remote: Total 17 (delta 4), reused 0 (delta 0), pack-reused 0
Receiving objects: 100% (17/17), 12.12 KiB | 1.10 MiB/s, done.
Resolving deltas: 100% (4/4), done.
(base) PS C:\Users\kris_\python> cd easy-local-rag
(base) PS C:\Users\kris_\python\easy-local-rag> pip install -r requirements.txt

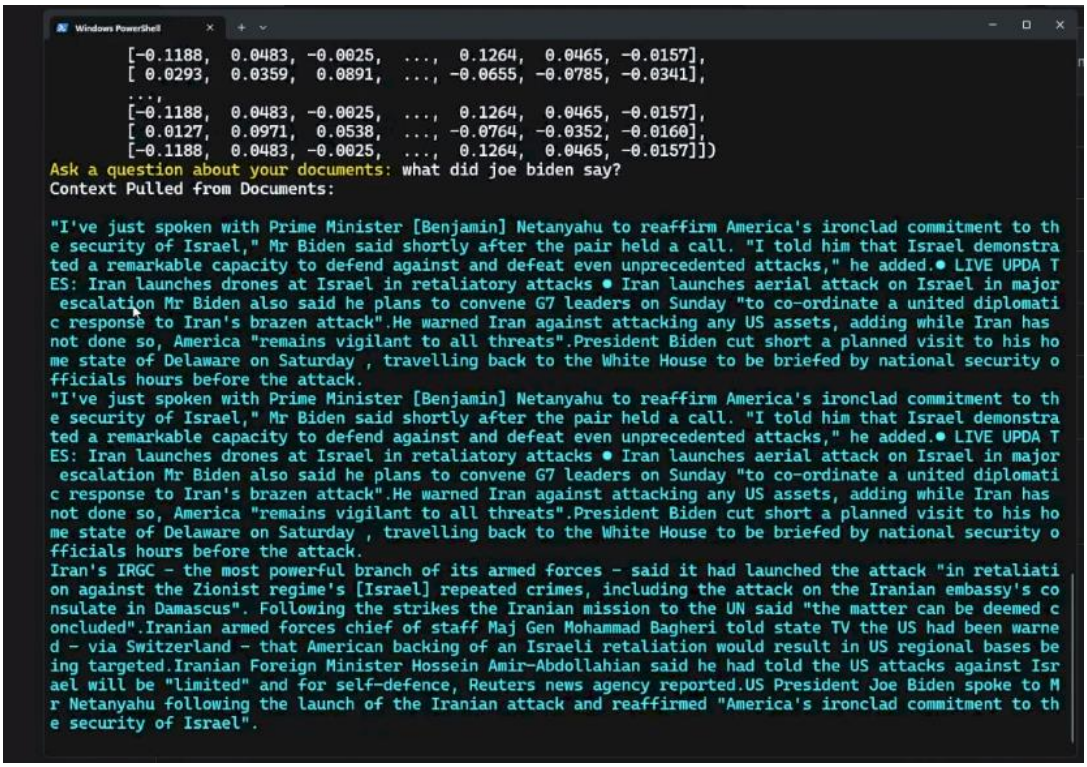

re<0.18.0,>=0.15.0->httpx<1,>=0.23.0->openai->-r requirements.txt (line 2)) (0.12.0)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\kris_\appdata\roaming\python\python311
\site-packages (from requests->huggingface-hub>=0.15.1->sentence-transformers->-r requirements.txt (line 1
)) (3.3.2)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\kris_\miniconda3\lib\site-packages (from req
uests->huggingface-hub>=0.15.1->sentence-transformers->-r requirements.txt (line 1)) (2.2.1)
WARNING: Skipping C:\Users\kris_\miniconda3\Lib\site-packages\urllib3-1.26.18.dist-info due to invalid met
adata entry 'name'
[notice] A new release of pip is available: 23.3.1 -> 24.0
[notice] To update, run: python.exe -m pip install --upgrade pip
(base) PS C:\Users\kris_\python\easy-local-rag> python pdf.py

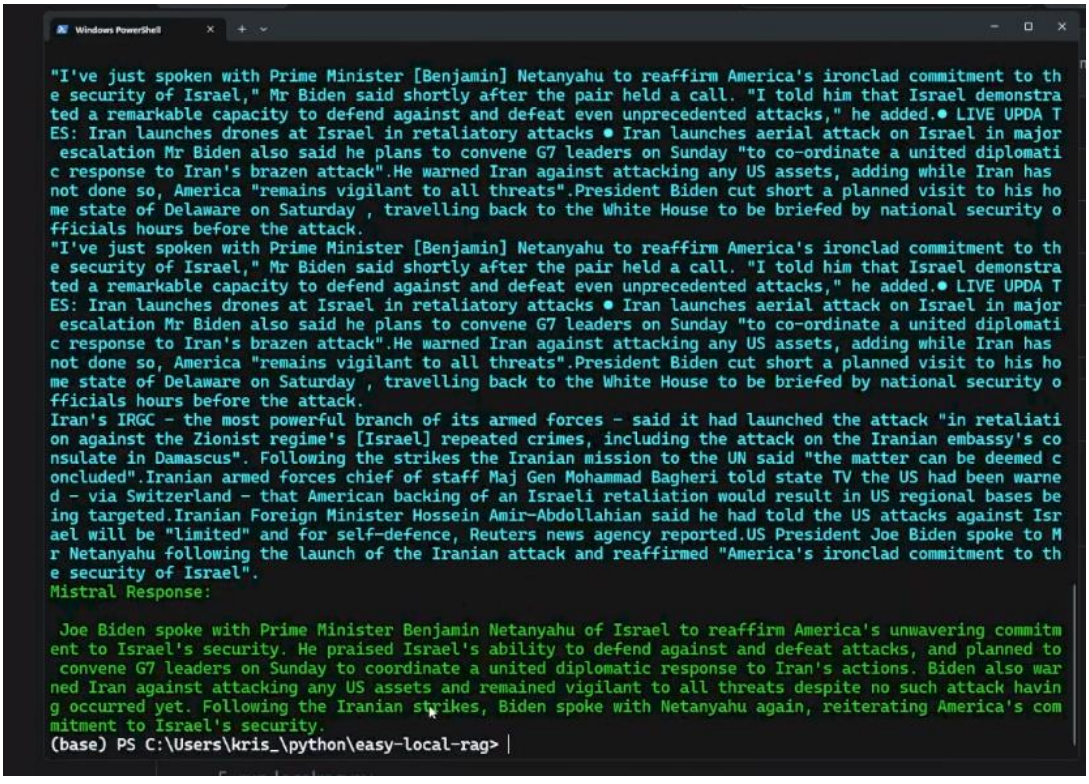



```
\site-packages (from requests->huggingface-hub>=0.15.1->sentence-transformers->-r requirements.txt (line 1
)) (3.3.2)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\kris_\miniconda3\lib\site-packages (from req
uests->huggingface-hub>=0.15.1->sentence-transformers->-r requirements.txt (line 1)) (2.2.1)
WARNING: Skipping C:\Users\kris_\miniconda3\Lib\site-packages\urllib3-1.26.18.dist-info due to invalid met
adata entry 'name'

[notice] A new release of pip is available: 23.3.1 -> 24.0
[notice] To update, run: python.exe -m pip install --upgrade pip
(base) PS C:\Users\kris_\python\easy-local-rag> python pdf.py
PDF content appended to vault.txt with each chunk on a separate line.
```

```
)) (3.3.2)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\kris_\miniconda3\lib\site-packages (from req
uests->huggingface-hub>=0.15.1->sentence-transformers->-r requirements.txt (line 1)) (2.2.1)
WARNING: Skipping C:\Users\kris_\miniconda3\Lib\site-packages\urllib3-1.26.18.dist-info due to invalid met
adata entry 'name'

[notice] A new release of pip is available: 23.3.1 -> 24.0
[notice] To update, run: python.exe -m pip install --upgrade pip
(base) PS C:\Users\kris_\python\easy-local-rag> python pdf.py
PDF content appended to vault.txt with each chunk on a separate line.
Traceback (most recent call last):
  File "C:\Users\kris_\python\easy-local-rag\pdf.py", line 53, in <module>
    root.mainloop()
  File "C:\Users\kris_\miniconda3\Lib\tkinter\__init__.py", line 1485, in mainloop
    self.tk.mainloop(n)
KeyboardInterrupt
(base) PS C:\Users\kris_\python\easy-local-rag> python localrag.py
Embeddings for each line in the vault:
tensor([[ 0.0546,  0.0915,  0.0460,  ..., -0.0449, -0.0332, -0.0185],
        [-0.1188,  0.0483, -0.0025,  ...,  0.1264,  0.0465, -0.0157],
        [ 0.0293,  0.0359,  0.0891,  ..., -0.0655, -0.0785, -0.0341],
        ...,
        [-0.1188,  0.0483, -0.0025,  ...,  0.1264,  0.0465, -0.0157],
        [ 0.0127,  0.0971,  0.0538,  ..., -0.0764, -0.0352, -0.0160],
        [-0.1188,  0.0483, -0.0025,  ...,  0.1264,  0.0465, -0.0157]])
Ask a question about your documents:
```

5. run localrag.py


Windows PowerShell
[-0.1188, 0.0483, -0.0025, ..., 0.1264, 0.0465, -0.0157],
[0.0293, 0.0359, 0.0891, ..., -0.0655, -0.0785, -0.0341],
...,
[-0.1188, 0.0483, -0.0025, ..., 0.1264, 0.0465, -0.0157],
[0.0127, 0.0971, 0.0538, ..., -0.0764, -0.0352, -0.0160],
[-0.1188, 0.0483, -0.0025, ..., 0.1264, 0.0465, -0.0157]])
Ask a question about your documents: what did joe biden say?
Context Pulled from Documents:
"I've just spoken with Prime Minister [Benjamin] Netanyahu to reaffirm America's ironclad commitment to th e security of Israel," Mr Biden said shortly after the pair held a call. "I told him that Israel demonstra ted a remarkable capacity to defend against and defeat even unprecedented attacks," he added.● LIVE UPDA T ES: Iran launches drones at Israel in retaliatory attacks ● Iran launches aerial attack on Israel in major escalation Mr Biden also said he plans to convene G7 leaders on Sunday "to co-ordinate a united diplomati c response to Iran's brazen attack".He warned Iran against attacking any US assets, adding while Iran has not done so, America "remains vigilant to all threats".President Biden cut short a planned visit to his ho me state of Delaware on Saturday , travelling back to the White House to be briefed by national security o fficials hours before the attack.
"I've just spoken with Prime Minister [Benjamin] Netanyahu to reaffirm America's ironclad commitment to th e security of Israel," Mr Biden said shortly after the pair held a call. "I told him that Israel demonstra ted a remarkable capacity to defend against and defeat even unprecedented attacks," he added.● LIVE UPDA T ES: Iran launches drones at Israel in retaliatory attacks ● Iran launches aerial attack on Israel in major escalation Mr Biden also said he plans to convene G7 leaders on Sunday "to co-ordinate a united diplomati c response to Iran's brazen attack".He warned Iran against attacking any US assets, adding while Iran has not done so, America "remains vigilant to all threats".President Biden cut short a planned visit to his ho me state of Delaware on Saturday , travelling back to the White House to be briefed by national security o fficials hours before the attack.
Iran's IRGC - the most powerful branch of its armed forces - said it had launched the attack "in retaliati on against the Zionist regime's [Israel] repeated crimes, including the attack on the Iranian embassy's co nsulate in Damascus". Following the strikes the Iranian mission to the UN said "the matter can be deemed c oncluded".Iranian armed forces chief of staff Maj Gen Mohammad Bagheri told state TV the US had been warne d - via Switzerland - that American backing of an Israeli retaliation would result in US regional bases be ing targeted.Iranian Foreign Minister Hossein Amir-Abdollahian said he had told the US attacks against Isr ael will be "limited" and for self-defence, Reuters news agency reported.US President Joe Biden spoke to M r Netanyahu following the launch of the Iranian attack and reaffirmed "America's ironclad commitment to th e security of Israel".


Windows PowerShell
"I've just spoken with Prime Minister [Benjamin] Netanyahu to reaffirm America's ironclad commitment to th e security of Israel," Mr Biden said shortly after the pair held a call. "I told him that Israel demonstra ted a remarkable capacity to defend against and defeat even unprecedented attacks," he added.● LIVE UPDA T ES: Iran launches drones at Israel in retaliatory attacks ● Iran launches aerial attack on Israel in major escalation Mr Biden also said he plans to convene G7 leaders on Sunday "to co-ordinate a united diplomati c response to Iran's brazen attack".He warned Iran against attacking any US assets, adding while Iran has not done so, America "remains vigilant to all threats".President Biden cut short a planned visit to his ho me state of Delaware on Saturday , travelling back to the White House to be briefed by national security o fficials hours before the attack.
"I've just spoken with Prime Minister [Benjamin] Netanyahu to reaffirm America's ironclad commitment to th e security of Israel," Mr Biden said shortly after the pair held a call. "I told him that Israel demonstra ted a remarkable capacity to defend against and defeat even unprecedented attacks," he added.● LIVE UPDA T ES: Iran launches drones at Israel in retaliatory attacks ● Iran launches aerial attack on Israel in major escalation Mr Biden also said he plans to convene G7 leaders on Sunday "to co-ordinate a united diplomati c response to Iran's brazen attack".He warned Iran against attacking any US assets, adding while Iran has not done so, America "remains vigilant to all threats".President Biden cut short a planned visit to his ho me state of Delaware on Saturday , travelling back to the White House to be briefed by national security o fficials hours before the attack.
Iran's IRGC - the most powerful branch of its armed forces - said it had launched the attack "in retaliati on against the Zionist regime's [Israel] repeated crimes, including the attack on the Iranian embassy's co nsulate in Damascus". Following the strikes the Iranian mission to the UN said "the matter can be deemed c oncluded".Iranian armed forces chief of staff Maj Gen Mohammad Bagheri told state TV the US had been warne d - via Switzerland - that American backing of an Israeli retaliation would result in US regional bases be ing targeted.Iranian Foreign Minister Hossein Amir-Abdollahian said he had told the US attacks against Isr ael will be "limited" and for self-defence, Reuters news agency reported.US President Joe Biden spoke to M r Netanyahu following the launch of the Iranian attack and reaffirmed "America's ironclad commitment to th e security of Israel".
Mistral Response:
Joe Biden spoke with Prime Minister Benjamin Netanyahu of Israel to reaffirm America's unwavering commitm ent to Israel's security. He praised Israel's ability to defend against and defeat attacks, and planned to convene G7 leaders on Sunday to coordinate a united diplomatic response to Iran's actions. Biden also war ned Iran against attacking any US assets and remained vigilant to all threats despite no such attack havin g occurred yet. Following the Iranian strikes, Biden spoke with Netanyahu again, reiterating America's com mitment to Israel's security.
(base) PS C:\Users\kris_\python\easy-local-rag>

More Agents Is All You Need

Junyou Li [*1] Qin Zhang [*1] Yangbin Yu [1] Qiang Fu [1] Deheng Ye [1]

Abstract

We find that, simply via a sampling-and-voting method, the performance of large language models (LLMs) scales with the number of agents instantiated. Also, this method is orthogonal to existing complicated methods to further enhance LLMs, while the degree of enhancement is correlated to the task difficulty. We conduct comprehensive experiments on a wide range of LLM benchmarks to verify the presence of our finding, and to study the properties that can facilitate its occurrence. Our code is publicly available at: Git.

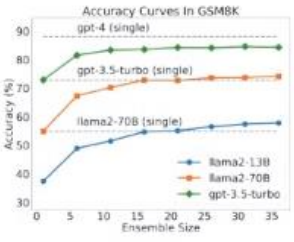


Figure 1. The accuracy increases with ensemble size across Llama2-13B, Llama2-70B and GPT-3.5-Turbo in GSM8K. When the ensemble size scales up to 15, Llama2-13B achieves comparable accuracy with Llama2-70B. Similarly, When the ensemble size scales up to 15 and 20, Llama2-70B and GPT-3.5-Turbo achieve comparable accuracy with their more powerful counterparts.

1. Introduction

Although large language models (LLMs) demonstrate remarkable capabilities in variety of applications (Zhao et al., 2023), such as language generation, understanding, and reasoning, they struggle to provide accurate answers when faced with complicated tasks. To improve the performance of LLMs, some of recent studies focus on ensemble methods (Wang et al., 2023b; Wan et al., 2024) and multiple LLM-Agents collaboration frameworks (Du et al., 2023; Wu et al., 2023).

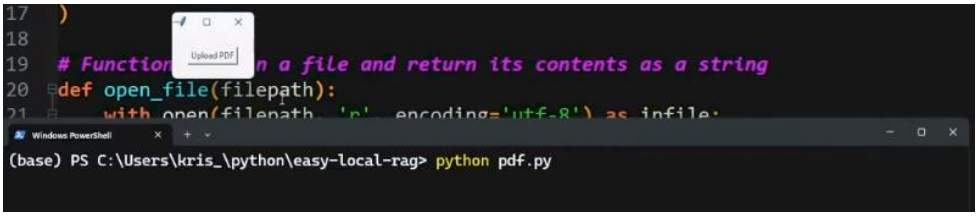
In these works, multiple LLM agents are used to improve the performance of LLMs. For instance, LLM-Debate (Du et al., 2023) employs multiple LLM agents in a debate form. The reasoning performance is improved by creating a framework that allows more than one agent to "debate" the final answer of arithmetic tasks. They show performance improvements compared to using one single agent. Similarly, CoT-SC (Wang et al., 2023b) generates multiple thought chains and picks the most self-consistent one as the final answer. The reasoning performance is improved by involving more thought chains compared to chain-of-thought (CoT) (Wei et al., 2022) which employs a single thought chain. Incidentally, from the data analysis of these works, we can notice the effects of putting multiple agents together, to some extent, can lead to a performance improvement in certain

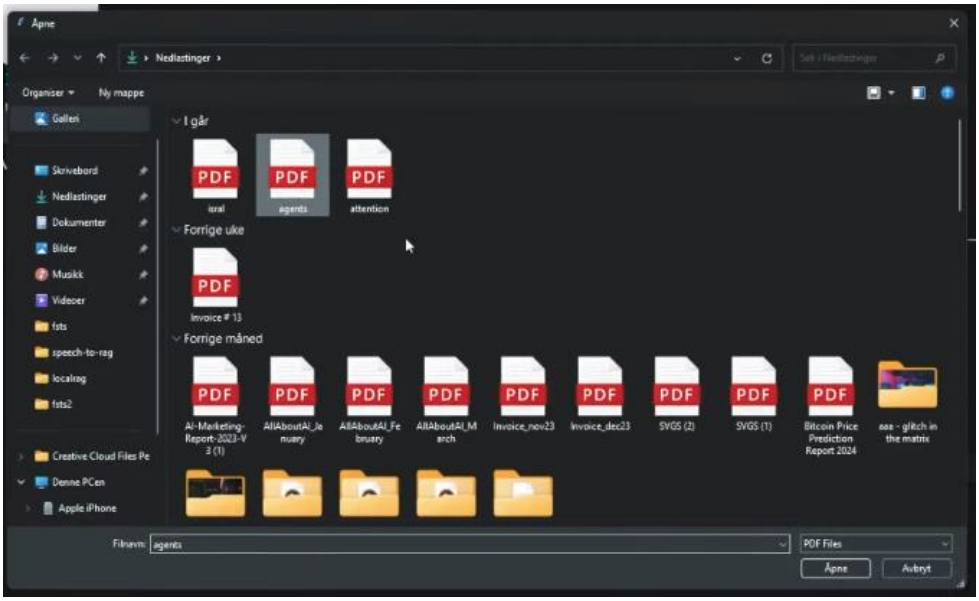
Debate (Du et al., 2023), the authors have reported a preliminary curve: the accuracy of a math problem increases with the number of debating agents (although the number was simply increased from 1 to 7). Also, in Wang et al. (2023b), involving more chains-of-thought pipelines (termed as a "sample-and-marginalize" decoding procedure), can lead to a performance gain. We realize that the LLM performance may likely be improved by a brute-force scaling up the number of agents instantiated. However, since the scaling property of "raw" agents is not the focus of these works, the scenarios/tasks and experiments considered are limited. So far, there lacks a dedicated in-depth study on this phenomenon. Hence, a natural question arises: *Does this phenomenon generally exist?*

To answer the research question above, we conduct the first comprehensive study on the *scaling property* of LLM agents. To dig out the potential of multiple agents, we propose to use a simple(st) sampling-and-voting method, which involves two phases. First, the query of the task, i.e., the input to

```python
import torch
from sentence_transformers import SentenceTransformer, util
import os
from openai import OpenAI

# ANSI escape codes for colors
PINK = '\033[95m'
CYAN = '\033[96m'
YELLOW = '\033[93m'
NEON_GREEN = '\033[92m'
RESET_COLOR = '\033[0m'

# Configuration for the Ollama API client
client = OpenAI(
    base_url='http://localhost:11434/v1',
    api_key='mistral'
)

# Function to open a file and return its contents as a string
def open_file(filepath):
    with open(filepath, 'r', encoding='utf-8') as infile:
        return infile.read()

# Function to get relevant context from the vault based on user input
def get_relevant_context(user_input, vault_embeddings, vault_content, model, top_k=3):
    if vault_embeddings.nelement() == 0:  # Check if the tensor has any elements
        return []
    # Encode the user input
    input_embedding = model.encode([user_input])
    # Compute cosine similarity between the input and vault embeddings
    cos_scores = util.cos_sim(input_embedding, vault_embeddings)[0]
    # Adjust top_k if it's greater than the number of available scores
    top_k = min(top_k, len(cos_scores))
    # Sort the scores and get the top-k indices
    top_indices = torch.topk(cos_scores, k=top_k)[1].tolist()
    # Get the corresponding context from the vault
    relevant_context = [vault_content[idx].strip() for idx in top_indices]
    return relevant_context


# Function to interact with the Ollama model
def ollama_chat(user_input, system_message, vault_embeddings, vault_content, model):
    # Get relevant context from the vault
    relevant_context = get_relevant_context(user_input, vault_embeddings, vault_content, model)
    if relevant_context:
        # Convert list to a single string with newlines between items
        context_str = "\n".join(relevant_context)
```

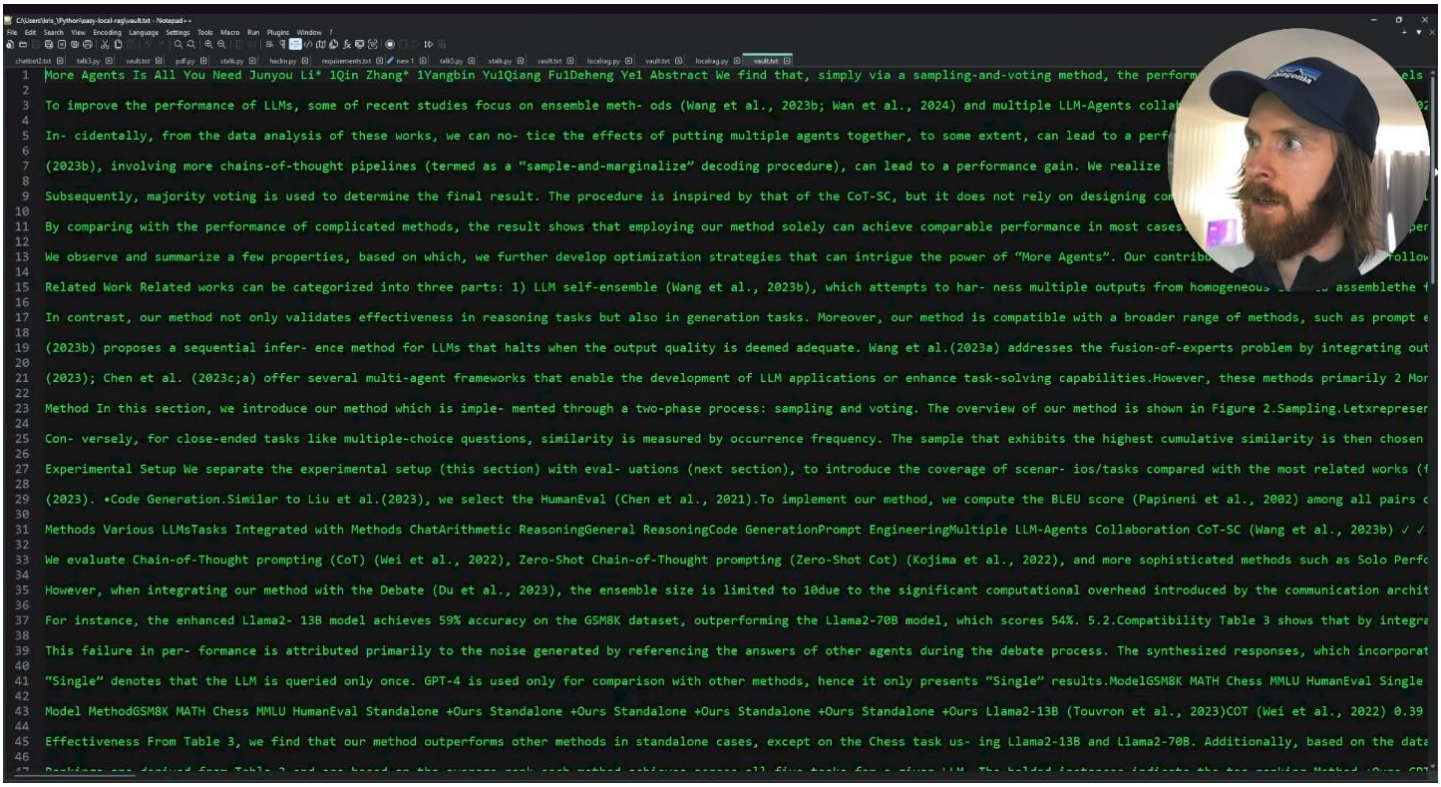
We can select the top 5 result as above



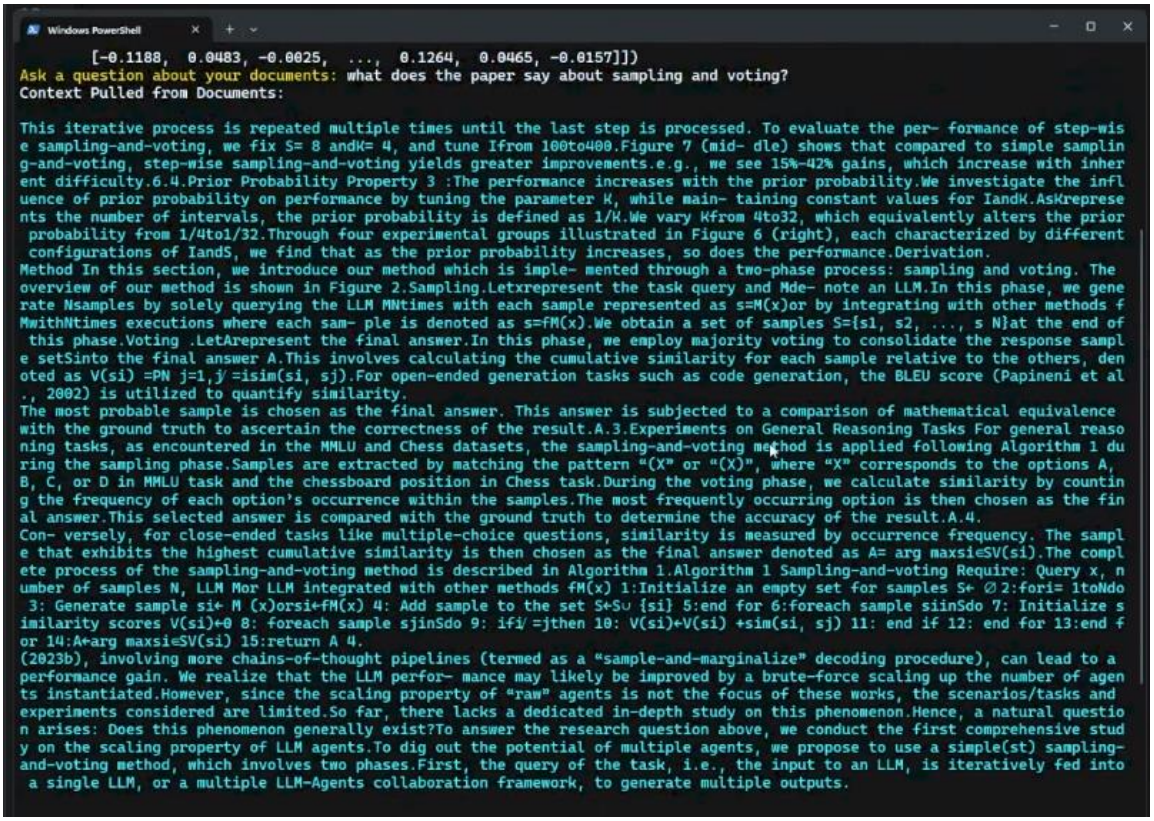


```
Windows PowerShell
(base) PS C:\Users\kris_\python\easy-local-rag> python pdf.py
PDF content appended to vault.txt with each chunk on a separate line.
```

```
Windows PowerShell
(base) PS C:\Users\kris_\python\easy-local-rag> python pdf.py
PDF content appended to vault.txt with each chunk on a separate line.
Traceback (most recent call last):
  File "C:\Users\kris_\python\easy-local-rag\pdf.py", line 53, in <module>
    root.mainloop()
  File "C:\Users\kris_\miniconda3\Lib\tkinter\__init__.py", line 1485, in mainloop
    self.tk.mainloop(n)
KeyboardInterrupt
(base) PS C:\Users\kris_\python\easy-local-rag>
```




```
Windows PowerShell
(base) PS C:\Users\kris_\python\easy-local-rag> python pdf.py
PDF content appended to vault.txt with each chunk on a separate line.
Traceback (most recent call last):
  File "C:\Users\kris_\python\easy-local-rag\pdf.py", line 53, in <module>
    root.mainloop()
  File "C:\Users\kris_\miniconda3\Lib\tkinter\__init__.py", line 1485, in mainloop
    self.tk.mainloop(n)
KeyboardInterrupt
(base) PS C:\Users\kris_\python\easy-local-rag> python localrag.py
Embeddings for each line in the vault:
tensor([[ 0.0171, -0.0677,  0.0357,  ...,  0.0657, -0.0240, -0.0251],
        [-0.1188,  0.0483, -0.0025,  ...,  0.1264,  0.0465, -0.0157],
        [ 0.0147, -0.0356, -0.0624,  ..., -0.0162, -0.0235,  0.0074],
        ...,
        [-0.1188,  0.0483, -0.0025,  ...,  0.1264,  0.0465, -0.0157],
        [ 0.0583, -0.0165, -0.1196,  ..., -0.0911, -0.1165, -0.0325],
        [-0.1188,  0.0483, -0.0025,  ...,  0.1264,  0.0465, -0.0157]])
Ask a question about your documents: what does the paper say about sampling and voting?
```



Now we have 5 chunks back to send to the LLM