Book Summarizer and Q&A chatbot

Arya Sinha, Ashutosh Dubey, Devyani Koshal, Mohammed Kaif,
Rohan Kulkarni, Shantanu Dixit

Group no: 34

Updated baseline

Baseline Model: Meta Llama-2-7B.

Previously, our approach involved fine-tuning the model on the downstream task of book summarization using the dataset BookSum. However, books often exceed the maximum token limit and fine-tuning has proven to be computationally and memory-intensive, particularly when applied to the task of book summarization.

With RAG, we extract relevant content, ensuring that it automatically conforms to the model's token restrictions, thereby eliminating the necessity for fine-tuning in this scenario.

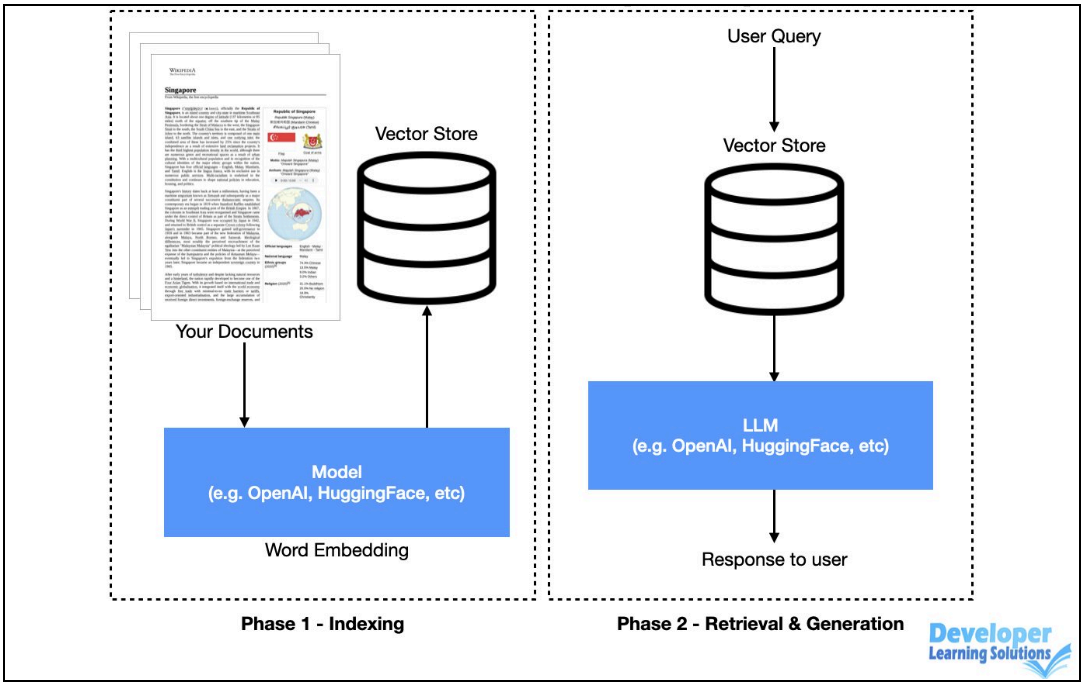
Proposed method

Retrieval-Augmented Generation combines the strengths of retrieval-based methods for information gathering with generation-based approaches for text synthesis. By integrating retrieval mechanisms with generation models, RAG systems can produce text that is grounded in factual information, contextually relevant, and coherent.

It majorly consists of two steps:

1. Retrieval: RAG starts by retrieving relevant information from a corpus of documents. The retrieval process involves matching the input query or context with the contents of the knowledge source to identify passages or documents that contain relevant information.

2. Generation: The retrieved information, along with any additional context or user input, is fed into a generation model. The generation model synthesizes the retrieved content with the input context to produce relevant text.



Phases of RAG pipeline [source]

We have implemented the retrieval phase of the RAG pipeline in the following way:

1. Document Loading

- Documents are retrieved from a specific directory, which contain various files containing textual data relevant to the domain or task at hand.
- In RAG, the documents serve as the corpus from which relevant information will be retrieved and used to generate responses to user queries.

2. Embedding Model

- An embedding model is initialized, specifically designed for generating embeddings from text inputs. These embeddings capture the semantic meaning of the text.

3. Service Context Creation

- A service context is established, bundling commonly used resources essential for indexing and querying documents within the RAG framework. This includes parameters like chunk size, language model configurations, and embedding model specifications.
- The service context ensures that all necessary resources are readily available during the retrieval and generation process, facilitating efficient information processing and response generation.

4. Vector Store Indexing

- The documents retrieved initially are indexed into a vector store index. This indexing process involves converting each document into a numerical vector representation using embeddings.
- These vectors are then stored in the index, enabling efficient retrieval of relevant documents during the retrieval phase of RAG. The indexed documents are the basis for generating informative responses to user queries.

Implementation details [Github repo]

Parameters Used

- *Context Window Size: 4096 tokens.*
- *Maximum New Tokens: 256 tokens.*
- *Generation Settings:*
  - *Temperature: 0.0 (no randomness).*
  - *Sampling: Disabled.*
- *Model and Tokenizer Names: "meta-llama/Llama-2-7b-chat-hf".*
- *Device Mapping: Auto selection.*
- *Model Precision: torch.float16 with 8-bit loading enabled (if supported).*
- *Chunk Size: 1024 tokens.*

Model Used:

- *Large Language Model (LLM):*
  - *Model Name: "meta-llama/Llama-2-7b-chat-hf".*
  - *Tokenizer Name: "meta-llama/Llama-2-7b-chat-hf".*

Embedding Model:

- *Model Name: "sentence-transformers/all-mpnet-base-v2".*

Initialization Details:

- *Document Loading: PDF documents are loaded from a specified directory.*
- *System Prompt: A system prompt provides guidance for the Q&A assistant.*
- *Query Wrapper Prompt: A template for user query input is established.*
- *Authentication: Token authentication is performed to access external resources.*

Service Context:

- Chunk Size: 1024 tokens.
- LLM Instance: Large Language Model (LLM) instance is included.
- Embedding Model Instance: Embedding model instance is included.

____
__

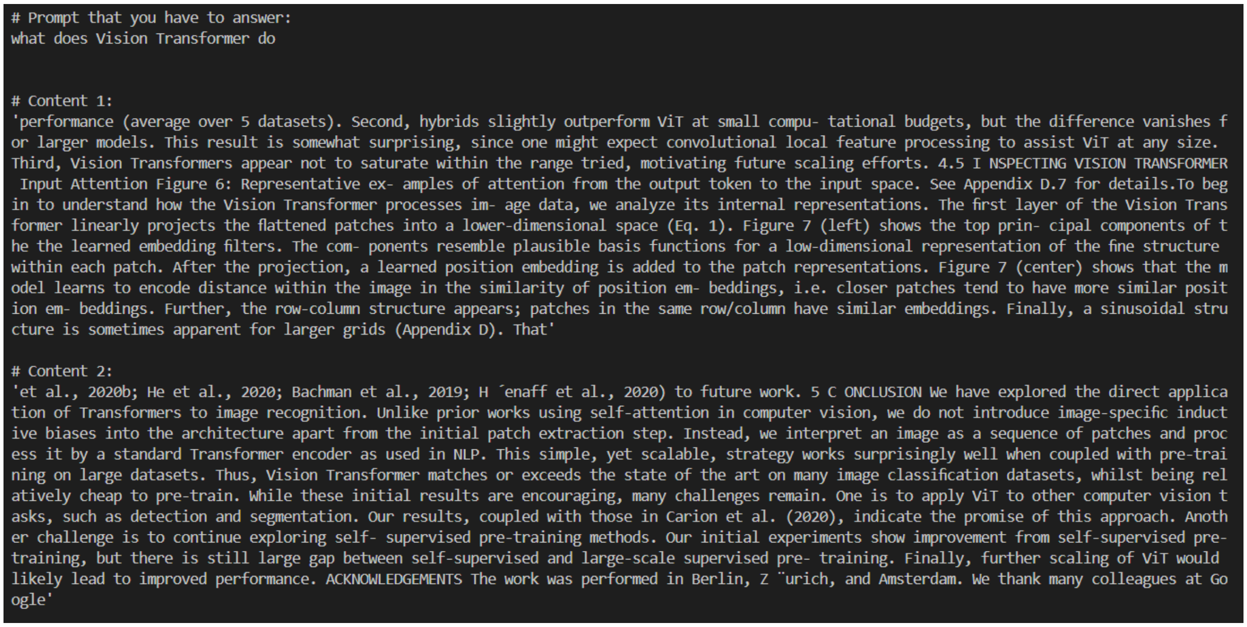
Vector Store Indexing:
- Index Creation: Documents are indexed into a vector store index.
- Model Usage: Embeddings generated from "sentence-transformers/all-mpnet-base-v2" are used for efficient document retrieval.

____
__

Results

The following snippets demonstrates the extracted relevant information from the document corpus corresponding to a user query.

Query 1

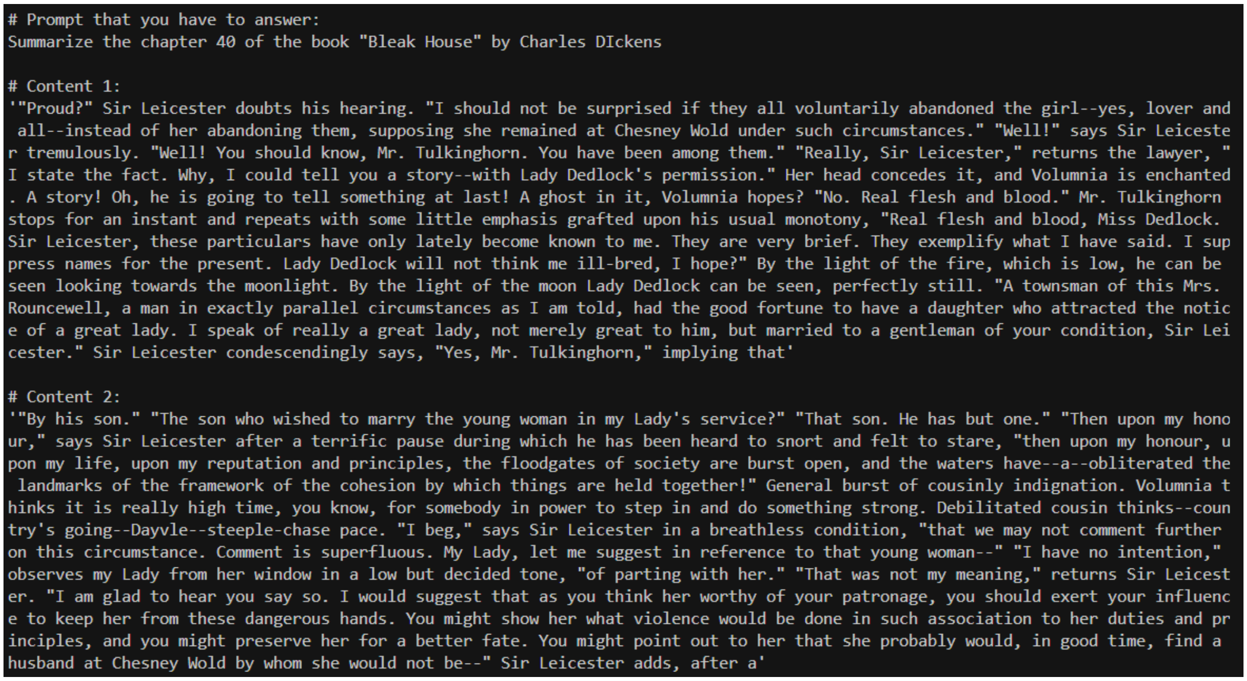


```
# Prompt that you have to answer:
what does Vision Transformer do


# Content 1:
'performance (average over 5 datasets). Second, hybrids slightly outperform ViT at small compu- tational budgets, but the difference vanishes f
or larger models. This result is somewhat surprising, since one might expect convolutional local feature processing to assist ViT at any size.
Third, Vision Transformers appear not to saturate within the range tried, motivating future scaling efforts. 4.5 I NSPECTING VISION TRANSFORMER
 Input Attention Figure 6: Representative ex- amples of attention from the output token to the input space. See Appendix D.7 for details.To beg
in to understand how the Vision Transformer processes im- age data, we analyze its internal representations. The first layer of the Vision Trans
former linearly projects the flattened patches into a lower-dimensional space (Eq. 1). Figure 7 (left) shows the top prin- cipal components of t
he the learned embedding filters. The com- ponents resemble plausible basis functions for a low-dimensional representation of the fine structure
within each patch. After the projection, a learned position embedding is added to the patch representations. Figure 7 (center) shows that the m
odel learns to encode distance within the image in the similarity of position em- beddings, i.e. closer patches tend to have more similar posit
ion em- beddings. Further, the row-column structure appears; patches in the same row/column have similar embeddings. Finally, a sinusoidal stru
cture is sometimes apparent for larger grids (Appendix D). That'

# Content 2:
'et al., 2020b; He et al., 2020; Bachman et al., 2019; H ´enaff et al., 2020) to future work. 5 C ONCLUSION We have explored the direct applica
tion of Transformers to image recognition. Unlike prior works using self-attention in computer vision, we do not introduce image-specific induct
ive biases into the architecture apart from the initial patch extraction step. Instead, we interpret an image as a sequence of patches and proc
ess it by a standard Transformer encoder as used in NLP. This simple, yet scalable, strategy works surprisingly well when coupled with pre-trai
ning on large datasets. Thus, Vision Transformer matches or exceeds the state of the art on many image classification datasets, whilst being rel
atively cheap to pre-train. While these initial results are encouraging, many challenges remain. One is to apply ViT to other computer vision t
asks, such as detection and segmentation. Our results, coupled with those in Carion et al. (2020), indicate the promise of this approach. Anoth
er challenge is to continue exploring self- supervised pre-training methods. Our initial experiments show improvement from self-supervised pre-
training, but there is still large gap between self-supervised and large-scale supervised pre- training. Finally, further scaling of ViT would
likely lead to improved performance. ACKNOWLEDGEMENTS The work was performed in Berlin, Z ¨urich, and Amsterdam. We thank many colleagues at Go
ogle'
```

Query 2

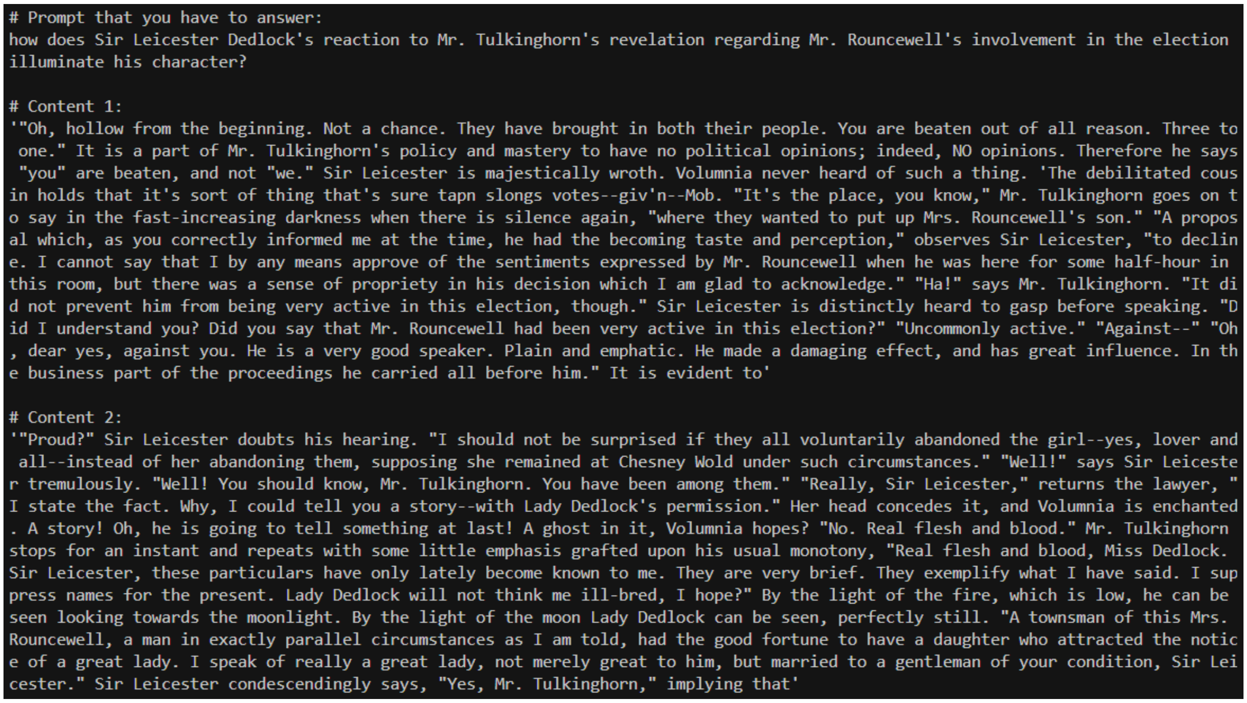


```
# Prompt that you have to answer:
Summarize the chapter 40 of the book "Bleak House" by Charles DIckens

# Content 1:
'"Proud?" Sir Leicester doubts his hearing. "I should not be surprised if they all voluntarily abandoned the girl--yes, lover and
 all--instead of her abandoning them, supposing she remained at Chesney Wold under such circumstances." "Well!" says Sir Leiceste
r tremulously. "Well! You should know, Mr. Tulkinghorn. You have been among them." "Really, Sir Leicester," returns the lawyer, "
I state the fact. Why, I could tell you a story--with Lady Dedlock's permission." Her head concedes it, and Volumnia is enchanted
. A story! Oh, he is going to tell something at last! A ghost in it, Volumnia hopes? "No. Real flesh and blood." Mr. Tulkinghorn
stops for an instant and repeats with some little emphasis grafted upon his usual monotony, "Real flesh and blood, Miss Dedlock.
Sir Leicester, these particulars have only lately become known to me. They are very brief. They exemplify what I have said. I sup
press names for the present. Lady Dedlock will not think me ill-bred, I hope?" By the light of the fire, which is low, he can be
seen looking towards the moonlight. By the light of the moon Lady Dedlock can be seen, perfectly still. "A townsman of this Mrs.
Rouncewell, a man in exactly parallel circumstances as I am told, had the good fortune to have a daughter who attracted the notic
e of a great lady. I speak of really a great lady, not merely great to him, but married to a gentleman of your condition, Sir Lei
cester." Sir Leicester condescendingly says, "Yes, Mr. Tulkinghorn," implying that'

# Content 2:
'"By his son." "The son who wished to marry the young woman in my Lady's service?" "That son. He has but one." "Then upon my hono
ur," says Sir Leicester after a terrific pause during which he has been heard to snort and felt to stare, "then upon my honour, u
pon my life, upon my reputation and principles, the floodgates of society are burst open, and the waters have--a--obliterated the
 landmarks of the framework of the cohesion by which things are held together!" General burst of cousinly indignation. Volumnia t
hinks it is really high time, you know, for somebody in power to step in and do something strong. Debilitated cousin thinks--coun
try's going--Dayvle--steeple-chase pace. "I beg," says Sir Leicester in a breathless condition, "that we may not comment further
on this circumstance. Comment is superfluous. My Lady, let me suggest in reference to that young woman--" "I have no intention,"
observes my Lady from her window in a low but decided tone, "of parting with her." "That was not my meaning," returns Sir Leicest
er. "I am glad to hear you say so. I would suggest that as you think her worthy of your patronage, you should exert your influenc
e to keep her from these dangerous hands. You might show her what violence would be done in such association to her duties and pr
inciples, and you might preserve her for a better fate. You might point out to her that she probably would, in good time, find a
husband at Chesney Wold by whom she would not be--" Sir Leicester adds, after a'
```

Query 3



```
# Prompt that you have to answer:
how does Sir Leicester Dedlock's reaction to Mr. Tulkinghorn's revelation regarding Mr. Rouncewell's involvement in the election
illuminate his character?

# Content 1:
'"Oh, hollow from the beginning. Not a chance. They have brought in both their people. You are beaten out of all reason. Three to
 one." It is a part of Mr. Tulkinghorn's policy and mastery to have no political opinions; indeed, NO opinions. Therefore he says
 "you" are beaten, and not "we." Sir Leicester is majestically wroth. Volumnia never heard of such a thing. 'The debilitated cous
in holds that it's sort of thing that's sure tapn slongs votes--giv'n--Mob. "It's the place, you know," Mr. Tulkinghorn goes on t
o say in the fast-increasing darkness when there is silence again, "where they wanted to put up Mrs. Rouncewell's son." "A propos
al which, as you correctly informed me at the time, he had the becoming taste and perception," observes Sir Leicester, "to declin
e. I cannot say that I by any means approve of the sentiments expressed by Mr. Rouncewell when he was here for some half-hour in
this room, but there was a sense of propriety in his decision which I am glad to acknowledge." "Ha!" says Mr. Tulkinghorn. "It di
d not prevent him from being very active in this election, though." Sir Leicester is distinctly heard to gasp before speaking. "D
id I understand you? Did you say that Mr. Rouncewell had been very active in this election?" "Uncommonly active." "Against--" "Oh
, dear yes, against you. He is a very good speaker. Plain and emphatic. He made a damaging effect, and has great influence. In th
e business part of the proceedings he carried all before him." It is evident to'

# Content 2:
'"Proud?" Sir Leicester doubts his hearing. "I should not be surprised if they all voluntarily abandoned the girl--yes, lover and
 all--instead of her abandoning them, supposing she remained at Chesney Wold under such circumstances." "Well!" says Sir Leiceste
r tremulously. "Well! You should know, Mr. Tulkinghorn. You have been among them." "Really, Sir Leicester," returns the lawyer, "
I state the fact. Why, I could tell you a story--with Lady Dedlock's permission." Her head concedes it, and Volumnia is enchanted
. A story! Oh, he is going to tell something at last! A ghost in it, Volumnia hopes? "No. Real flesh and blood." Mr. Tulkinghorn
stops for an instant and repeats with some little emphasis grafted upon his usual monotony, "Real flesh and blood, Miss Dedlock.
Sir Leicester, these particulars have only lately become known to me. They are very brief. They exemplify what I have said. I sup
press names for the present. Lady Dedlock will not think me ill-bred, I hope?" By the light of the fire, which is low, he can be
seen looking towards the moonlight. By the light of the moon Lady Dedlock can be seen, perfectly still. "A townsman of this Mrs.
Rouncewell, a man in exactly parallel circumstances as I am told, had the good fortune to have a daughter who attracted the notic
e of a great lady. I speak of really a great lady, not merely great to him, but married to a gentleman of your condition, Sir Lei
cester." Sir Leicester condescendingly says, "Yes, Mr. Tulkinghorn," implying that'
```

References

1. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D., 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
2. https://www.youtube.com/watch?v=f-AXdiCyiT8&t=7s
3. https://www.youtube.com/watch?v=1FERFfut4Uw&t=3176s