Query My Library! Question-Answering with Open-Source LLMs and Local Books

Razvan Florian Vasile [1]

[1]Computer Science, University of Bologna

Introduction

Suppose an avid reader with an extensive bookshelf struggles to fully grasp challenging ideas in complex texts. Could we facilitate their learning by integrating an assistant to help them study? I explore the use of Transformers that leverage **agent planning** pipelines and **RAG engines** for a hybrid approach using both **supervised** and **unsupervised learning** techniques.

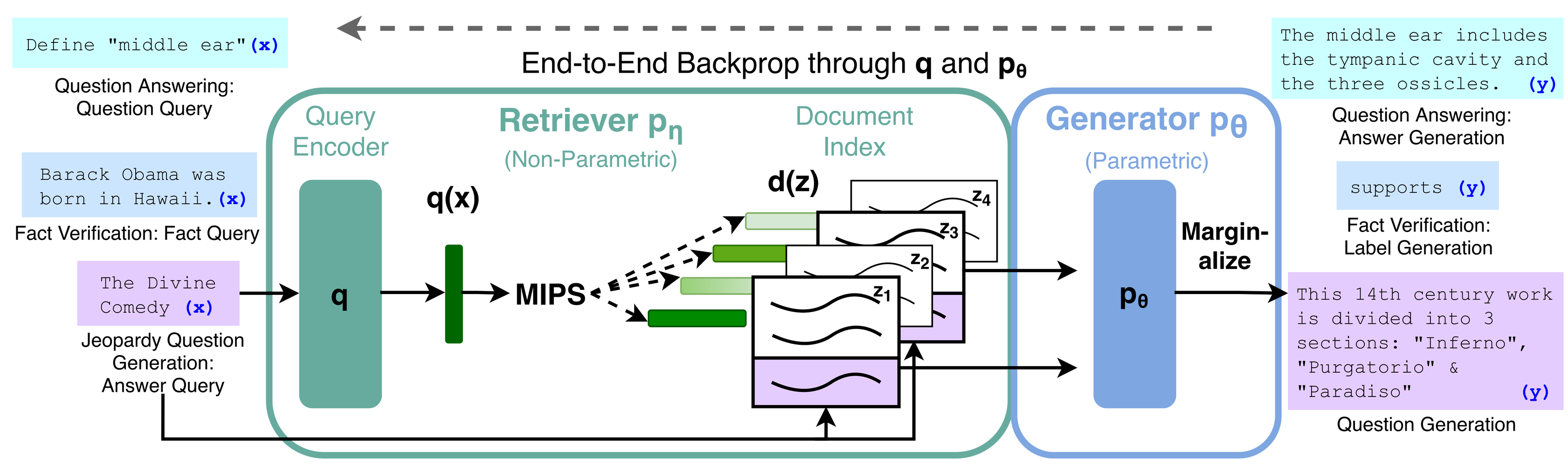


Figure 1. The RAG Architecture. Taken from [3].

Problem Statement

- **Identify the Problem:** Traditional LLMs struggle with knowledge-intensive tasks, such as specialized domain questions, fact verification, or Jeopardy-style questions.
- **Put it into context:** Answering questions about books may exhibit these limitations.
- **Find the root cause:** The model often fails to locate precise book passages, leading to incomplete or inaccurate answers.
- **Ideal outcome:** Provide the model with relevant information to generate accurate responses.
- **Propose a solution:** Develop a method to expand prompts with essential information and iteratively refine answers for accuracy.

Results: Generated Responses with Citations

The Little Prince by Antoine de Saint-Exupéry

What is the significance of the Little Prince's relationship with his home planet and his responsibility towards it?

Thought: I need to use a tool to help me answer the question.
Action: LittlePrince
Action Input: ["Input: "Little Prince's relationship with his home planet."]

Source 1. [...] When a serious reason to believe that the planet from which the little prince came is the asteroid known as B-612. This asteroid has only once been seen through the telescope.

Source 2. The planet was tiny. Over what could this king really rule? [...] "Sire, over what do you rule? "Over everything, said the king, with magnificent simplicity. Such power was a thing for the little prince to marvel at. If he had been master of such complete authority, he would have been able to watch the sunset, not forty-four times in one day, but seventy-two, or even a hundred, or even two hundred times, without ever having to move his chair. And because he felt a bit sad as he remembered his little planet, which he had forsaken.

Source 3. "What is this 'where I live,' of which you speak? Where do you want to take your sheep? [...] But the little prince seemed shocked by this offer: "Tie him! What a queer idea!" "But if you don't tie him, I said, "he will wander off somewhere, and get lost. My friend broke into another peal of laughter: "But where do you think he would go?" "Anywhere. Straight ahead of him."

Observation. The Little Prince's relationship with his home planet is one of deep connection and care. The Little Prince's responsibility towards his home planet is also hinted at through his conversation with the king. When the king boasts about his absolute and universal rule, the Little Prince is amazed but also seems to be thinking about his own role as a caretaker of his small planet. [...]

Answer. The significance of the Little Prince's relationship with his home planet lies in its representation of a deep emotional bond, care, and responsibility. This connection is reflected in his concern for the sheep he left behind, his appreciation for comfort and familiarity, and his feelings of loss when remembering it. His conversation with the king also suggests that he takes his role as a caretaker seriously, valuing the well-being of his planet and its inhabitants.

Ethics by Baruch Spinoza

What does Spinoza mean by "intellectual love of God" (), and how does this concept function in his ethical system?

Thought: I need to use a tool to help me answer the question.
Action: Ethics
Action Input: ["Input: "What does Spinoza mean by "intellectual love of God"?]

Source 1. [...] This joy is purely intellectual, encompassing the joy of knowing one's own essence perfectly through that of God. This is what Spinoza calls 'intellectual love of God' (P32C). This is not like love of another person, love of a thing of even the love of God of P15.Intellectual love of God is knowing and loving all being eternally (P33)

Source 2. [...] Those adequate ideas are also accompanied by active joy, which means that the free person understands God to be the cause of his joy [...]. The more ade-quate ideas we have, the stronger is the idea of God and the stronger is our love for God. Our love of God grows with our true understanding. [...] our idea of God is necessarily adequate and cannot be accompanied by any kind of sadness [...]

Source 3. [...] God's love of being is better understood as the affirmation of God's perfection. And it is the affirmation of the perfection of ourselves, in God, that constitutes our blessedness (P365). [...]

Observation. Spinoza's "intellectual love of God" refers to a purely intellectual joy that arises from knowing one's own essence perfectly through that of God. This concept is not comparable to the love of another person, thing, or even the love of God in other contexts.
In Spinoza's ethical system, this concept functions as a central virtue. It grows with our true understanding, and since our idea of God is necessarily adequate, it cannot be accompanied by sadness or hate.

Answer. Spinoza's concept of "intellectual love of God" (amor intellectualis dei) refers to a profound understanding and joy that arises from knowing one's own essence in relation to God's essence. This concept is central to his ethical system, where it grows with our true understanding and constitutes our blessedness. It is considered the most constant of affects and the greatest virtue, as it involves rejoicing in all being eternally.

Methods: How is the Vector Store generated?

- **Book Set:** We have N books, labeled as $b_1, b_2, \ldots, b_N$.
- **Page Segmentation:** Each book b_i is divided into pages. Let's call the k-th page of book i as $p_{i,k}$.
- **Semantic Chunking:** For each page $p_{i,k}$, we create multiple chunks of text based on semantic meaning. Let's call these chunks $c_{i,k,1}, c_{i,k,2}, \ldots$
  We then use an embedding function f (as described in [4]) to convert each chunk into a 384-dimensional vector: $e_{i,k,l} = f(c_{i,k,l})$. Here, $e_{i,k,l}$ is the embedding of the l-th chunk from the k-th page of the i-th book.
- **Vector Database:** We collect all these embeddings to create a vector database D, as described in [1].
- **User Query:** When a user submits a query q, we:
  1. Convert the query to an embedding using the same function: $e_q = f(q)$.
  2. Use an SVM to find the top 3 closest matches in our database D.

Why use two embedding models?

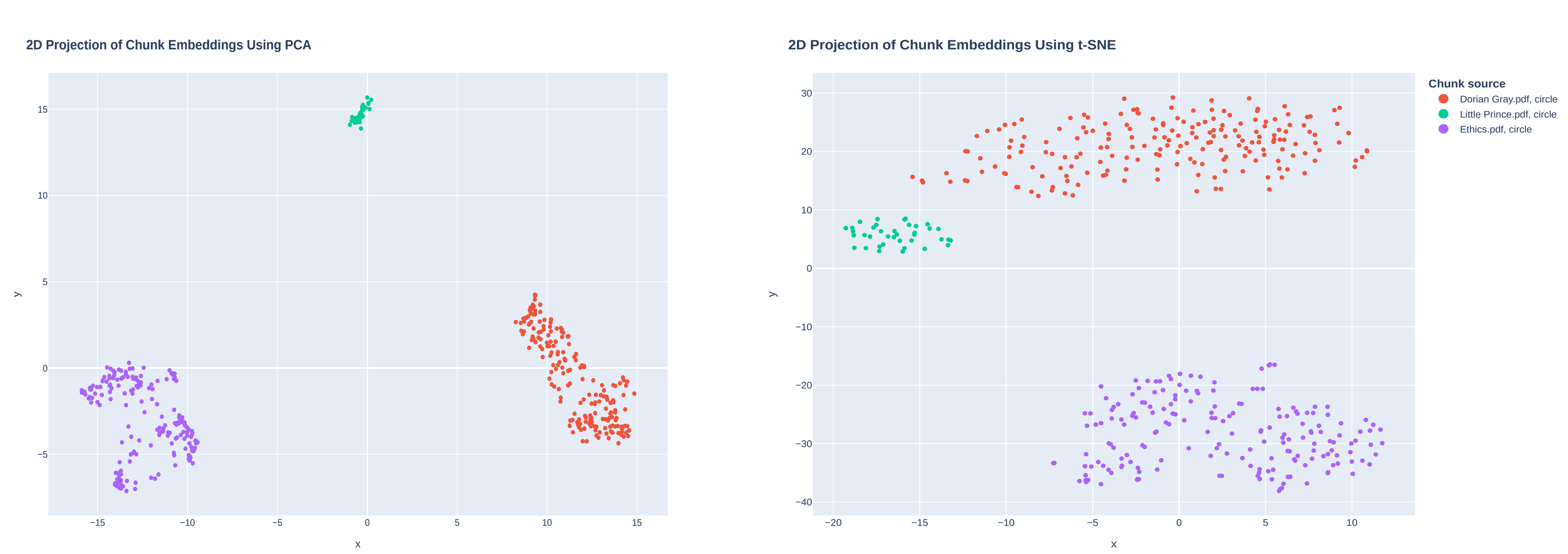
- **Query Embedding [1].** Small (33M parameters) yet powerful model optimized for applications where low latency is a requirement.
- **Semantic Chunking [4].** Comparatively, larger model (109M parameters) designed for semantic search. It seeks to improve search accuracy by understanding the semantic meaning of the search query and the corpus to search over.

Model Name	Use case	Dimensions	Size	Suitable Score Function
BAAI/bge-small-en-v1.5 [1]	querying	384	33M	SVM [2]
multi-qa-mpnet-base-dot-v1 [4]	semantic nodes	768	109M	dot-product

Table 1. Properties of the embedding models used to create semantic nodes and to generate the vector store.

Analysis: 2-Dimensional Cluster Visualization

- **PCA.** Is a linear dimension reduction technique that seeks to maximize variance and preserves large pairwise distances. In other words, items that are different end up far apart.
- **t-SNE.** It differs from PCA by preserving only small pairwise distances or local similarities. Gives an intuition on how data is arranged in higher dimensions.



(a) PCA 2D Projection (b) t-SNE 2D Projection

Figure 2. Three clusters: The Picture of Dorian Gray (red), The Little Prince (green), Ethics (purple).

Pairwise Heatmap Embeddings

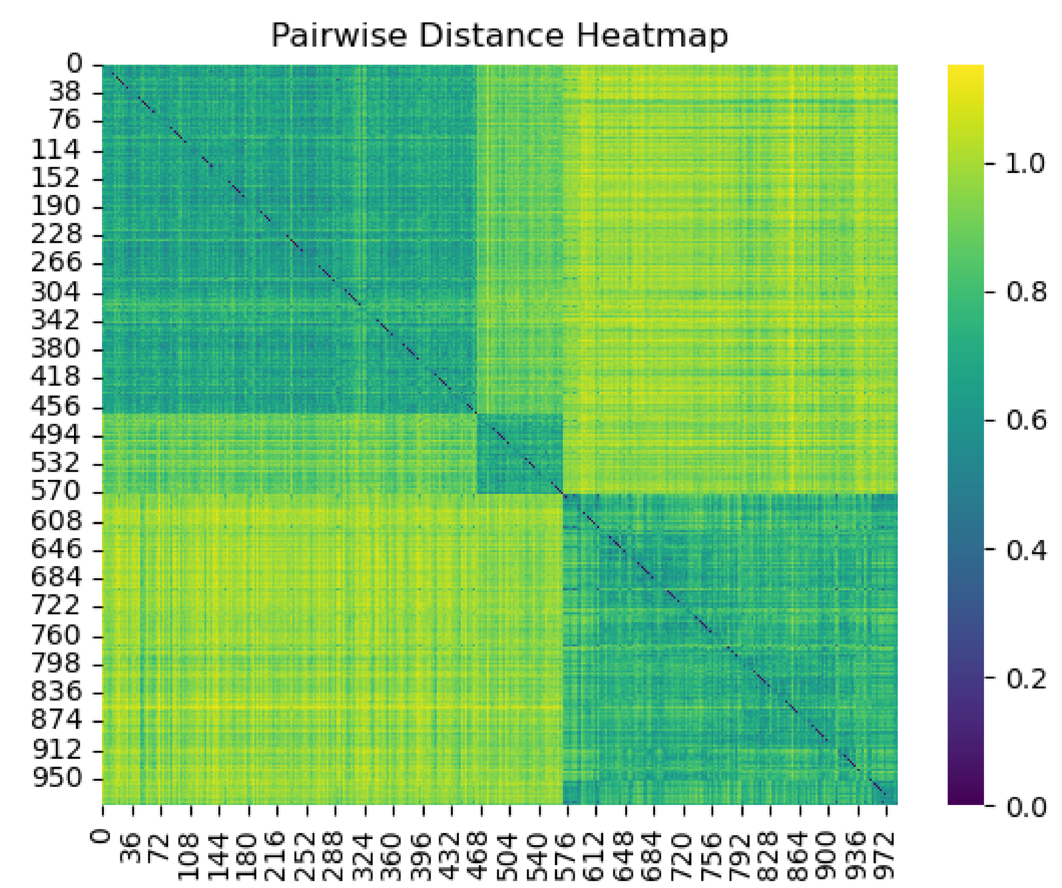


Figure 3. Upper left corner: The Picture of Dorian Gray, Middle: The Little Prince, Bottom Right: Ethics.

- **Dark colors** represent lower pairwise distances, indicating **more similar pairings**.
- **Lighter colors** represent higher pairwise distances, indicating **less similar pairings**.

Conclusion

- **Local and remote inference.** The model generally requires strong LLMs such as *Llama 3* in order to adequately use complex planning pipelines. Fortunately, Llama 3 is available for free both locally and remotely through the HuggingFace API.
- **Future models.** Due to the fact that the Librarian is usable both locally and via remote inference, the project is likely to scale well as stronger open-source models become available.

Future Work

- **Integrate with other services.** Use other resources made available through the LlamaIndex hub, including the ability to search Wikipedia or to query the arXiv knowledge database.
- **Enable Multi-agent workflows.** Allow users to interact with multiple agents by asking and answering questions and engaging in interactive conversations[5].

References

[1] BAAI/bge-small-en-v1.5 · Hugging Face — huggingface.co.
`https://huggingface.co/BAAI/bge-small-en-v1.5`.

[2] Andrej Karpathy.
`https://github.com/karpathy/randomfun/blob/master/knn_vs_svm.ipynb`.

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela.
Retrieval-augmented generation for knowledge-intensive nlp tasks, 2020.

[4] sentence-transformers/multi-qa-mpnet-base-dot-v1 · Hugging Face — huggingface.co.
`https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1`.

[5] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang.
Autogen: Enabling next-gen llm applications via multi-agent conversation framework.

 razvanflorian.vasile@studio.unibo.it