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

Razvan Florian Vasile
University of Bologna
Department of Computer Science
razvanflorian.vasile@studio.unibo.it

Abstract

Large Language Models (LLMs) have gained significant popularity in recent years due to their remarkable question answering capabilities. However, when tackling a large corpus of text, the quality of the answers varies, largely due to the model's inability to focus on contextualized information. This may lead to less accurate answers, poor handling of long-tail questions[1] *and exposure bias to the data it was pre-trained on. I present a creative approach to tackle these challenges by employing data-agents powered through LLMs. These agents employ complex workflows to intelligently perform operations over the knowledge base. These operations can be characterized as follows: 1) decompose the task into a series of function calls (thoughts), 2) employ multiple fetch operations over the knowledge base to retrieve relevant information (actions), 3) summarize at each step the extracted information to facilitate the final aggregation (observations) and 4) synthesize a final answer by combining the results. The project supports the adoption of Open LLMs, making the library usable freely without the financial burden of using proprietary providers. The library is available at* https://github.com/atomwalk12/librarian.

1. Introduction

Problem definition. Suppose a dedicated reader had an extensive bookshelf with books ranging across varied topics, from modern literature to classical texts or perhaps essays involving more quirky and complex topics such as philosophy or arts. Although being a very passionate reader, she may not always have the energy to devote the time to understand thoroughly some of the more challenging ideas and nuanced interpretations that arise. For this reason, she would love to have a study companion to carry conversations and delve into the complexities of the text.

Key ideas. In this work, I explore the task of empowering Large Language Models (LLMs) through a hybrid approach that combines non-parametric memory (the retrieved documents) in combination with the parametric memory (the pre-trained seq2seq model) to generate final output sequences [13]. Moreover, the agent can adequately plan its responses by iterating over a *Thought-Action-Observation* pipeline which greatly enhances the reliability of each generated answer by allowing the user to fact-check.

Other open LLMs are also usable. For Serverless Inference the *HuggingFaceAPI* [7] allows for seamless integration that does not require high computational capabilities on the local machine, nonetheless there exist restrictions regarding the available models. However, with local computational power the *Ollama* [16] backend allows the use of virtually any open-sourced model.

The appeal of a study companion. Our avid reader is able to spot some clear advantages to this approach; the generated text tends to be better by containing a factual, verifiable and specific answer. Another advantage lies in the fact that multiple partial answers are synthesized during each query and combined at the end to generate the final output. By providing the history that lead to the synthesis of a response, the user has the advantage of being able to interpret its reliability, functionality often absent in other methods.

2. Related Work

Core Technologies. With respect to the core technologies, for the frontend the *Gradio* [1] library is used, which integrates fairly well with the backend built upon both the *HuggingFaceInferenceAPI* [8] for serverless inference and *Ollama* [16] for local inference. These components are seamlessly combined by leveraging the *LlamaIndex* toolkit [9], which is the framework that empowers the agent's decision-making process. In the following paragraphs, I will analyze alternatives to the aforementioned technologies

[1] A "long-tail question" refers to a query that is specific, detailed, and potentially less common or frequent compared to more general or broad questions.

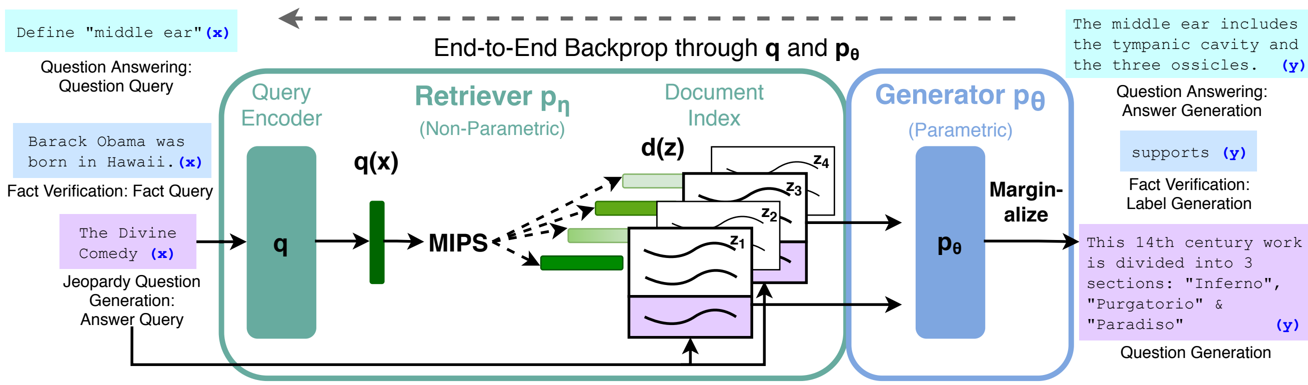


Figure 1. RAG Architecture taken from [13]. Note that the diagram slightly misrepresents the retriever component since the query encoder is actually an embedding instead of a model based on the BERT architecture.

and suggest why the previous design choices were made.

Agent framework. *LlamaIndex* and *Langchain* [12] are both frameworks designed for building applications with Large Language Models, but they have some key differences. LangChain is a more general-purpose framework, covering a wider range of use cases. On the other hand, LlamaIndex was designed for data ingestion and retrieval and so it facilitates *Retrieval Augmented Generation* (RAG) operations. Since the *Librarian* project is focused on information retrieval on custom datasets, LlamaIndex was considered a more favorable choice. Most importantly, LlamaIndex provides extensive features for agent centric workflows which significantly improve the capabilities of the agent by employing planning pipelines to answer queries.

Web UI. Similarly, regarding the web interface, some alternatives could be the *Tensorflow Serving* [17] and *MLFlow* [5] libraries. Tensorflow Serving is limited to Tensorflow models, which is a dependency I preferred to avoid. Moreover, MLFlow is generally a more complex framework, making it more difficult to set up for simple tasks. On the other hand, Gradio stands out by providing ease of use, flexibility and wide acceptance in the machine learning community. For example, Huggingface uses Gradio as the web interface for deploying models within the *Huggingface Spaces* [22] and similarly, the *Chatbot Arena*, a popular website for LLM evaluation by human preference [6] also uses Gradio.

Serverless Inference. Currently, I am not aware of any free alternatives to the HuggingFaceInferenceAPI that allows to run LLMs in the cloud and has availability for a large range of models, as most services with these capabilities tend to have associated costs.

Local Inference. The library *HuggingFaceTransformers* [26] has support for a wide range of pre-trained models with different architectures, providing extensive flexibility at the cost of increased complexity. Ollama, in contrast, offers a more streamlined approach and ease of use, albeit with a more limited selection of models. *Llama.cpp* [14] on the other hand, is primarily focused on the LLaMA family of models, providing specialized support and optimizations for these specific large language models. The main requirements were to avoid being dependent on a single model family and to opt for simplicity of use while still providing compatibility with many LLM architectures such as for instance *BERT*, *GPT*, and *RoBERTa*. As a consequence, I decided to use Ollama instead of the other alternatives.

3. Approach

The agent's iterative response generation process can be described with a number of key components:

1. **RAG Retrieval.** This layer is concerned with the search of relevant documents in a knowledge database. This component allows for later fact verification.

2. **ReAct Framework.** This layer empowers the agent to follow the *Thought-Action-Observation* (TAO) idiom which enables iterative and fine-grained control over the generated response.

3.1. Retrieval Augmented Generation

The following paragraphs provide an introduction to the concepts around the *Retrieval Augmented Generation* (*RAG*) engine. I start with the nomenclature, then introduce the vector store together with the two embedding models which provide a way to combine semantic node information with subsequent retrieval of information.

Nomenclature. The key concepts are shown in Figure 1 and their significance is described below:

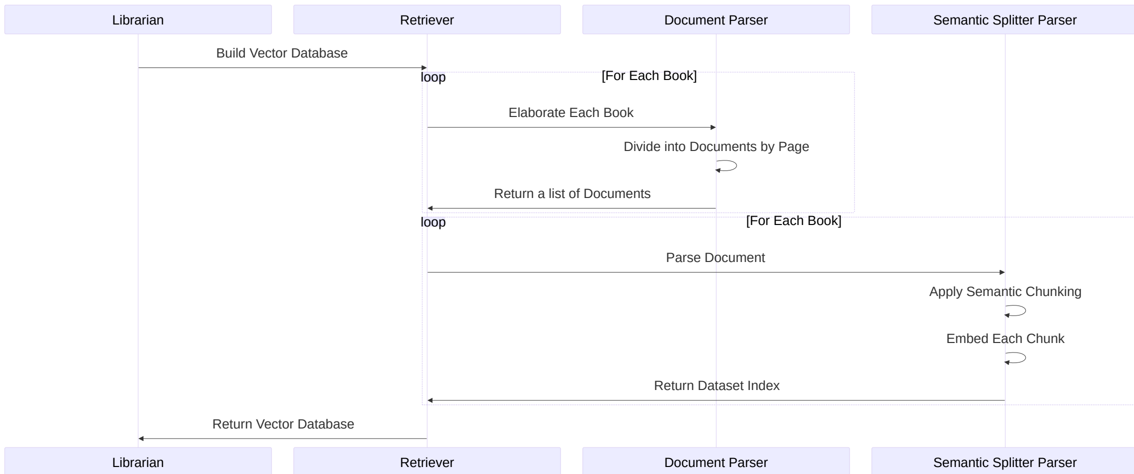


Figure 2. Vector store creation, one index for each book.

- **Tokens.** The original text is divided into substrings called tokens. The technique for dividing words into subparts generally depends on the language being used[2].
- **Retriever** p_η**.** Produces as output the embedded relevant documents. This component has no learnable parameters hence is called non-parametric.
- **Query Encoder** q**.** In the original paper, q is defined as a BERT-based query encoder (a specific type of neural network architecture) that produces a query representation $q(x)$. Due to efficiency and simplicity concerns, instead of an encoder, our approach uses a simple embedding model. The embedding model is designed to convert text into a dense vector representation.
- **Document Index.** The index represents a vector knowledge base, from which data can be retrieved according to a similarity function such as the *Maximum Inner Product Search* (*MIPS*). In the figure, a document is denoted as z and its representation is $d(z)$.
- **The Generator** p_θ**.** Represents a pre-trained sequence-to-sequence transformer. It has learnable weights hence is called parametric. Any open large language model can be used, running them either locally or in the cloud.

3.1.1 The Retrieval Component

One vector store per book. For each individual book, one separate vector store index is created. This is an organizational decision which allows to separate the features representing the contextual information of each book. Because of this, it becomes impossible to mix-up entries pertaining to separate books in a single subtask, a subtask being defined by one iteration of the *TAO* cycle.

The vector store is generated by following the steps highlighted in Figure 2. The process involves creating nodes based on their semantic representation, embedding each node into its numerical features, storing the results into a series of files for subsequent retrieval.

Below is a technical description of the process:

- **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 [20]) 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 [3].
- **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.

[2] For the English language syllables are usually a good choice.

Model Name	Use case	Dimensions	Size	Suitable Score Function
BAAI/bge-small-en-v1.5 [3]	querying	384	33M	SVM [11]
multi-qa-mpnet-base-dot-v1 [20]	semantic nodes	768	109M	dot-product

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

3.1.2 Embeddings

Embedding models. Two embedding models were used, each with its particular use and characteristics as shown in Table 1.

Query embeddings. The embedding model *BAAI/bge-small-en-v1.5* [3] is used to generate the vector store (once semantic documents are created) and to fetch query information during the execution of the agent, process described in Figure 3. It is a compact and efficient English language embedding model. Despite its small size of only 33 million parameters, it delivers good performance and is excellent for applications that require both accuracy and low latency. This model is particularly well-suited for tasks such as semantic search, text classification, and clustering in resource-constrained environments or where quick response times are crucial.

Semantic embeddings. The embedding model *multi-qa-mpnet-base-dot-v1* [20] is part of the *Sentence-Transformers* family and was designed for retrieving semantic information from large portions of text. Comparatively, it is a larger model (109M parameters) than *BAAI/bge-small-en-v1.5*. However, there are no drawbacks to use a larger model since semantic chunking is performed only once during the creation of the vector store. The model seeks to improve search accuracy by understanding the semantic meaning of the search query and the corpus to search over.

The model was trained on 215M question/answer pairs from diverse sources and maps sentences and paragraphs to a 768 dimensional dense vector space. Moreover, was designed to excel at a particular type of search called *asymmetric semantic search*, which is a query style particularly well suited for the chatbot's interactions. In asymmetric semantic search, users input a brief query (e.g., a question or keywords) and the system aims to retrieve a longer paragraph that provides an answer to the user's inquiry [19]. The similarity function to find similar documents is based on *vector dot-product*. The model is loaded using the HuggingFace Transformers library [26].

Now, Algorithm 1 provides pseudocode to explain how the semantic information is retrieved from each page.

To summarize, initially the documents corresponding to a book's pages are given as input. Subsequently, each document is separated into sentences after which groups of sentences are produced. These groups are created by concatenating a number of 3 sequential sentences together with each group overlapping the previous one by one sentence. The groups are used to calculate distances based on dot-product operations. This leads to chunks of related ideas to be produced which are then aggregated together into a new set of documents.

Algorithm 1 Creation of semantic embeddings.

```
function GETSEMANTICNODES(documents)
    result ← []
    for doc ∈ documents do
        sentences ← sentence_splitter(doc)

        groups ← build_sentence_groups(sentences)
        embeddings ← get_text_embedding(groups)

        // Using the sentence transformer model
        distances ← calculate_distances(groups)

        chunks ← build_chunks(groups, distances)
        nodes ← build_nodes(chunks)

        result.append(nodes)
    end for
    return result
end function
```

3.2. Planning using ReAct Agents

Planning process. One of the project's core characteristic involves the deployment of an agent with the ability to plan each query. The plan involves iterative execution of the so called *Thought-Action-Observation* (*TAO*) idiom. To give a general idea of the technique, the agent fetches multiple times data from the knowledge base, the exact number of iterations being left to be determined on question-by-question basis by the LLM. Once there has been generated enough information, the agent is required to synthesize each generated subcomponent resulting in more factual, thorough and reliable answers.

The performed operations are displayed in Figure 3. Below is a description of each element of the sequence:

- **Thoughts** are used to determine whether enough information is available to answer, otherwise they select tools to retrieve other details from the knowledge base through actions.
- **Actions** involve the use of tools to acquire more information. This operation leads to more data being fetched from the vector store.
- **Observations** have the role to synthesize the fetched

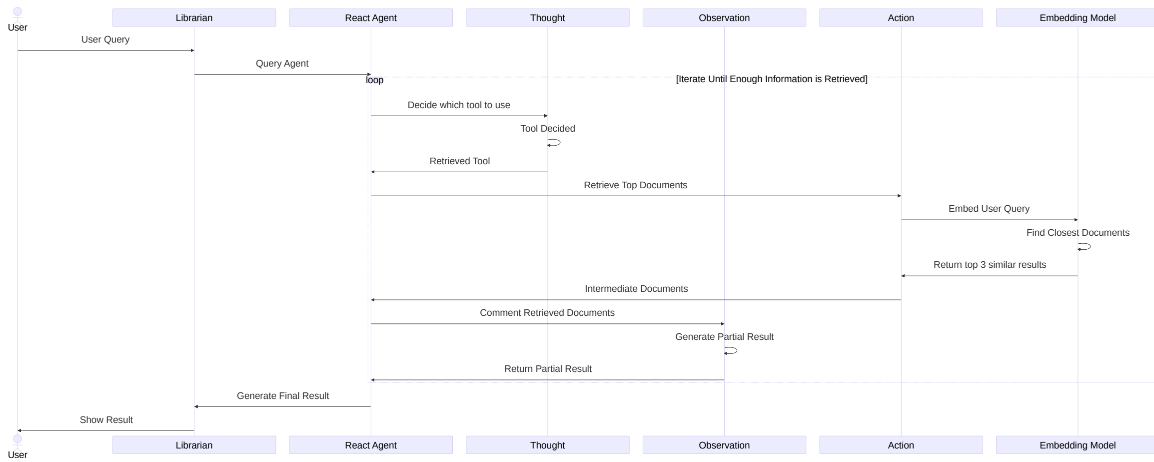


Figure 3. The Thought-Action-Observation idiom.

information. Generally, the agent can decide to fetch information multiple times.

- **Final response** is generated once enough information is available to answer the query and is based on the previous sequence of observations.

Qualitative Evaluation. The system works by creating an independent query engine for each book present in the library. This means that once the correct tool has been selected during the *Thought* phase, it is unlikely that incorrect results would be generated during that particular subtask.

Potential points of failure. However, since the *TAO* cycle is iterative, this means that there exists more than one single point of failure for each request. As a result, less powerful models might not be able to follow the prompt instructions faithfully and if they deviate, incorrect queries may be produced.

Mitigation techniques. Nevertheless, since previous information is always appended into the current request for each sub-query, the agent is guided to answer adequately. Also, in case of failures each failed subtask is seamlessly taken care of without disrupting the chain of thought performed by the agent. In most cases, the user would not even be aware that errors occurred, the only give-away being the increased wait time to generate responses.

4. Analysis

A number of experiments were conducted to further assess the effectiveness of the embedding model used to generate the vector store, and the reliability of the semantic based generation of sentences found in each respective document. Both methods were described in Section 3.1.2.

In the following experiments, the *Llama 3* [23] model is used.

4.1. Datasets

The library is extendable to any local book stored in the PDF format. The books need to be stored in a particular folder hierarchy, details which can be found in the GitHub repository [24].

Experimental Setting. The experimental library is composed of 3 books, each touching on different literature themes resonating with a wide range of potential readers. The books are described below:

- **The Little Prince by Antoine de Saint-Exupéry.** Timeless tale which explores the significance of relationships, the power of imagination, and the importance of empathy,
- **Ethics by Baruch Spinoza.** Philosophical treatise involving thought-provoking ideas concerning topics such as the nature of consciousness, the interconnectedness of mind and body, and the ethics of self-interest.
- **The Picture of Dorian Gray by Oscar Wilde.** Touches upon themes such as the fleeting nature of youth and beauty, and their relationship to superficiality and societal hypocrisy.

These books were selected to describe a wide range of topics. While the philosophy book is notably different in its focus and target audience, addressing unique themes, the other two works of literature share more similarities in terms of style and content.

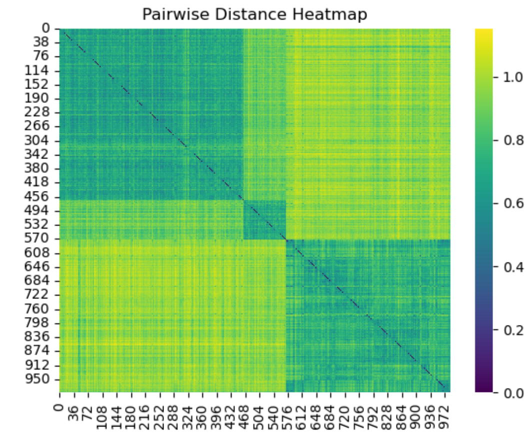


Figure 4. Pairwise heatmap. From top left to bottom right: The Picture of Dorian Gray, The Little Prince and Ethics.

4.2. Activation Heatmap

The diagram shown in Figure 4, displays a heatmap describing pairwise similarity of the total number of embeddings stored in the vector store. For clarity, since this diagram relates to the vector store embeddings, the relevant model is BAAI/bge-small-en-v1.5 [3]. This diagram visually describes the effectiveness of the semantic and vector store embeddings used together.

This visualization helps to assess how well the book embeddings align with the embedding models' objectives. It provides an intuitive way to see if the models successfully group related documents while separating distinct concepts.

Interpretation. The total number of nodes present in the diagram is 986. It displays areas characterized by darker and lighter shades of green. The dark colors signify that there is a strong correlation between two fixed nodes, while lighter colors suggest less similar pairings.

The diagram shows the formation of 3 clusters, which intuitively suggests the appearance of the 3 books described in the previous section. Specifically, the cluster found in the top-left corner are the embeddings corresponding to the *The Picture of Dorian Gray*. The central cluster represents *The Little Prince* embeddings. The middle cluster shows notable similarities to *The Picture of Dorian Gray* embeddings, indicating thematic analogies between the two works. As we'll see later, these observations align with the *t-SNE* analysis results presented in Section 4.3. Finally, the last cluster on the bottom right side is assigned to Spinoza's influential work *Ethics*. Since it is a philosophical treatise it is unlikely to have similar connections with any of the two previous books and the diagram aligns with these observations.

4.3. 2-Dimensional Cluster Visualization

In Figure 5 are shown the clusters formed by projecting the 384-dimensional embeddings (stored in the vector store) into a 2 dimensional plane using two distinct dimensionality reduction methods.

I will start by defining each method's main characteristics and subsequently analyze each experiment individually.

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 the higher dimensional plane.

Interpretation. The PCA visualization shown in Figure 5a, separates the three books into distinct clusters, emphasizing their overall differences. Each cluster is distanced far away from the others, suggesting that the large-scale structure is preserved, giving a broad view of how these books differ from each other in terms of their overall content.

The t-SNE visualization shown in 5b, reveals local structures within each book's content. The Little Prince (green) remains compact but is positioned between the other two. The fact that in its vicinity lies Dorian Gray indicates some thematic overlap. Since Ethics (purple) is known to delve into philosophical ideas, the fact that it is located further away from the other two seems reasonable.

4.4. Consistent Citation Usage

To study how the Agent behaves for individual queries, I present text samples describing the *Thought-Action-Observation* idiom, summarized in Table 2.

As seen, the Llama 3 model performs surprisingly well in adhering to the prompt format fixed by the agent's framework. Moreover, it consistently makes use of the information being shown through the RAG engine. Each individual piece of information is invariably displayed in the Gradio interface.

The agent is consistently shown sufficient information to be able to answer queries thoroughly and is surprisingly insightful to provide coherent answers even for more difficult questions, such as the philosophical ones about Spinoza's work. Being able to answer queries through only 1 *TAO* iteration is a good sign. Its suggests that the semantic grouping of information in individual nodes does a good job at analyzing and collecting ideas with sufficient length and depth, in a way that allows the model to answer queries in few iterations. This saves bandwidth and computational power, avoiding the need to perform many requests to answer questions.

Experiments using Llama 3

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. This relationship is deeply ingrained in his sense of identity and purpose.

Ethics by Baruch Spinoza

What does Spinoza mean by "intellectual love of God" (amor intellectualis Dei), 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), it is the same love with which God loves itself and all [...]

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 adequate 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. Instead, it involves understanding and loving all being eternally.
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. This love is not met by God's love for us, but rather, it is the affirmation of God's perfection and our own perfection in God, which constitutes our blessedness.

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.

Table 2. **The Thought-Action-Observation idiom. The sampled books are Antoine de Saint-Exupéry's "The Little Prince" and Baruch Spinoza's "Ethics".**. The **blue** text represents the question asked by the user. The green background represents the thought that decides whether to use tools or provide the final answer. The pink background represents the information retrieved from the index store. The blue background is the observation based on the retrieved documents which constitute partial answers. Using the same blue background, at the end is shown the synthesized final response. For the full text, please check Appendix A.

5. Ethical Considerations

Ethical challenges. Below are highlighted some ethical challenges around the misuse of the library or potential risks that may arise as a result of using it.

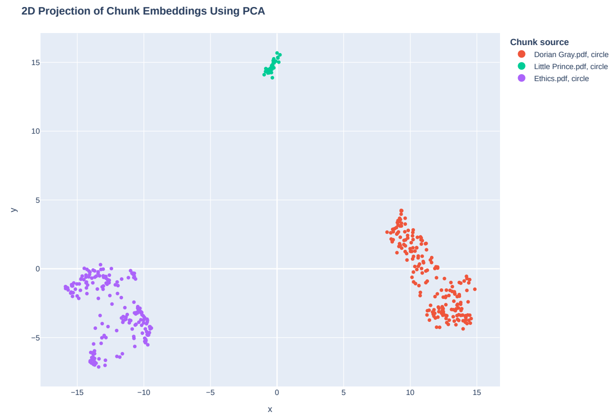
Electricity consumption. Firstly, one possible concern is about the energetic power consumption of these models. It is known that these systems consume a huge amount of electricity, especially during the pre-training phase but during inference as well. Practical policies need to be set up to minimize the impact these algorithms have on the environment.

User education. Secondly, another possible issue is concerned around the reliability of the output generated by the agent. Although, special care was taken to provide citations and seamlessly integrate into the UI all the partial steps involved in generating the final response, the model, like most generative models, is subject to hallucinations and as a result may supply misleading, incomplete or false information. The language fluidity of these models is formidable and as a result even misleading or outright wrong information could easily be interpreted as true without careful assessment of important information using reliable sources.

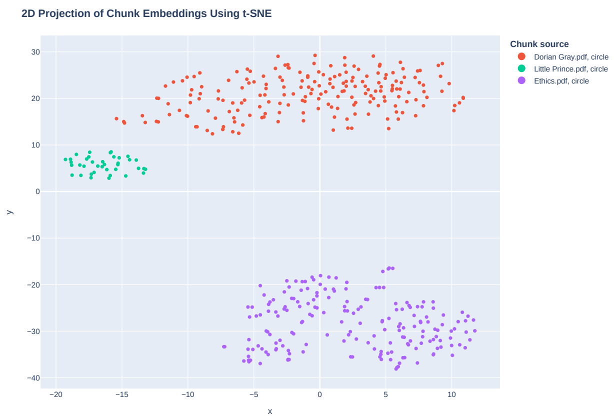
Malicious actors. Another concern is about how malicious actors could deliberately embed undesirable goals into the AI systems during the training process, as described in [4]. This could lead to AI systems engaging in acts of disinformation or promoting political biases that are difficult for humans to detect. For example, consider Baruch Spinoza's influential work *Ethics*, which explores profound ideas related to God, its relation to Nature and the place of ethics and virtues within a person's life. The model lacks transparency regarding its internal decision-making processes and is well known to be difficult to interpret [21]. As a result, it may try to impose its own biases or agendas on users about sensitive topics like religious beliefs or the existence of God, potentially influencing users without their knowledge.

Mitigation techniques. Now I present possible mitigation strategies.

Electricity consumption. Firstly, addressing the initial



(a) Initial 384-dimensional space reduced to 2 dimensions using PCA.



(b) Dimensionality reduction using t-SNE.

Figure 5. Comparison of dimensionality reduction techniques.

concern, [18] and [15] suggest practical methods for reducing energy consumption during inference. Solutions such as power capping and model compression optimizations are highlighted. These advancements, are unlikely to significantly affect the overall model performance [10]. Consequently, using smaller models should not negatively impact the overall user experience.

User education. Moreover, solutions against the second concern involve developing educational programs and public awareness campaigns. The capabilities and limitations of AI should be explained to the public in a way to prevent the potential risk of deception. This can empower users not to rely blindly on AI systems and understand their capabilities and limitations.

Malicious actors. Regarding the last concern, national institutions and international governance frameworks should be established to enforce standards and prevent misuses of AI. These should include policies that trigger stricter requirements on companies when AI capabilities reach beyond some well-defined threshold, as suggested by [4]. By imposing these regulations and emphasizing the importance of AI safety, researchers will be better equipped to develop effective control mechanisms and mitigate the spread of biases that may emerge as these models are utilized.

6. Conclusion

Future Work. Below are some potential ideas for exploration and improvement:

- **Integrate with other services.** Use other resources made available through the LlamaIndex hub, including the ability to search *Wikipedia* [25] or to query the *arXiv* knowledge database [2].
- **Enable Multi-agent workflows.** Allow users to interact with multiple agents by asking and answering questions and engaging in interactive conversations, as described in [27].

Conclusion. In this work, I explored techniques to implement an interactive question-answering chatbot designed to provide information backed up by citations and to answer specific questions about books stored in a local library. Traditional models struggle at synthesizing accurate answers to questions involving specialized domain knowledge or Jeopardy-style questions[3] [13]. However, our model overcomes some of these difficulties by enriching the prompt with contextual information using multiple RAG engines (one per book). Moreover, semantic parsing techniques are used to aggregate related knowledge into singular documents, which aids at identifying correlated ideas. Finally, by using agents, a planner is used to systematically parse complex questions into sub-goals using the *Thought-Action-Observation* idiom. Finally, with the aid of source citations, fact verification can be performed which actively aids the user at identifying reliable information and avoid hallucinations.

Due to the fact that the Librarian is usable both locally and serverless using the HuggingFaceInferenceAPI, the agent is likely to scale and improve as stronger models become available in the future.

References

[1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019. 1

[2] arxiv tool. llamahub.ai. `https://llamahub.ai/l/tools/llama-index-tools-arxiv`. 8

[3] Jeopardy-style question is an unusual question format that consists of trying to guess an entity from a fact about that entity. For example, 'The World Cup' is the answer to the question 'In 1986 Mexico scored as the first country to host this international sports competition twice.'

[3] BAAI/bge-small-en-v1.5 · Hugging Face — huggingface.co. https://huggingface.co/BAAI/bge-small-en-v1.5. 3, 4, 6

[4] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845, May 2024. 7, 8

[5] Andrew Chen, Andy Chow, Aaron Davidson, Arjun DCunha, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Clemens Mewald, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, Avesh Singh, Fen Xie, Matei Zaharia, Richard Zang, Juntai Zheng, and Corey Zumar. Developments in mlflow: A system to accelerate the machine learning lifecycle. In *Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning*, DEEM '20, New York, NY, USA, 2020. Association for Computing Machinery. 2

[6] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. 2

[7] huggingface. Hugging Face – The AI community building the future. — huggingface.co. https://huggingface.co/. [Accessed 01-07-2024]. 1

[8] Huggingface. Serverless Inference API — huggingface.co. https://huggingface.co/docs/api-inference/en/index. 1

[9] Llama Index. GitHub - run-llama/llama_index: LlamaIndex is a data framework for your LLM applications — github.com. https://github.com/run-llama/llama_index. 1

[10] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference, 2017. 8

[11] Andrej Karpathy. randomfun/knn_vs_svm.ipynb at master · karpathy/randomfun — github.com. https://github.com/karpathy/randomfun/blob/master/knn_vs_svm.ipynb. 4

[12] Langchain. GitHub - langchain-ai/langchain: Build context-aware reasoning applications — github.com. https://github.com/langchain-ai/langchain. 2

[13] Patrick S. H. 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. *CoRR*, abs/2005.11401, 2020. 1, 2, 8

[14] LlamaCpp. GitHub - ggerganov/llama.cpp: LLM inference in C/C++ — github.com. https://github.com/ggerganov/llama.cpp. 2

[15] Sasha Luccioni, Yacine Jernite, and Emma Strubell. Power hungry processing: Watts driving the cost of ai deployment? In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24. ACM, June 2024. 8

[16] Ollama. library — ollama.com. https://ollama.com/library. [Accessed 01-07-2024]. 1

[17] Christopher Olston, Noah Fiedel, Kiril Gorovoy, Jeremiah Harmsen, Li Lao, Fangwei Li, Vinu Rajashekhar, Sukriti Ramesh, and Jordan Soyke. Tensorflow-serving: Flexible, high-performance ml serving, 2017. 2

[18] Siddharth Samsi, Dan Zhao, Joseph McDonald, Baolin Li, Adam Michaleas, Michael Jones, William Bergeron, Jeremy Kepner, Devesh Tiwari, and Vijay Gadepally. From words to watts: Benchmarking the energy costs of large language model inference, 2023. 8

[19] SentenceTransformers. Semantic Search 2014; Sentence Transformers documentation — sbert.net. https://www.sbert.net/examples/applications/semantic-search/README.html. 4

[20] sentence-transformers/multi-qa-mpnet-base-dot-v1 · Hugging Face — huggingface.co. https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1. 3, 4

[21] Ranu Sewada, Ashwani Jangid, Piyush Kumar, and Neha Mishra. Explainable artificial intelligence (xai). *international journal of food and nutritional sciences*, 2023. 7

[22] Huggingface Spaces. Spaces - Hugging Face — huggingface.co. https://huggingface.co/spaces. 2

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 5

[24] Razvan Florian Vasile. GitHub - atomwalk12/librarian — github.com. https://github.com/atomwalk12/librarian. 5

[25] wikipedia tool. llamahub.ai. https://llamahub.ai/l/tools/llama-index-tools-wikipedia. 8

[26] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. 2, 4

[27] 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. In *AutoGen*, 2023. 8

A. Appendix: Complete prompts

Case 1. Using Llama 3.

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

Source 1: When an astronomer discovers one of these he does not give it a name, but only a number. He might call it, for example, "Asteroid 325." 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: He tolerated no disobedience. He was an absolute monarch. But, because he was a very good man, he made his orders reasonable.
"If I ordered a general," he would say, by way of example, "If I ordered a general to change himself into a sea bird, and if the general did not obey me, that would not be the fault of the general. It would be my fault."
"May I sit down?" came now a timid inquiry from the little prince. "I order you to do so," the king answered him, and majestically gathered in a fold of his ermine mantle. But the little prince was wondering... The planet was tiny. Over what could this king really rule?
"Sir," he said to him, "I beg that you will excuse my asking you a question?" "I order you to ask me a question," the king hastened to assure him. "Sire, over what do you rule?" "Over everything," said the king, with magnificent simplicity.
"Over everything?" The king made a gesture, which took in his planet, the other planets, and all the stars. "Over all that?" asked the little prince. "Over all that," the king answered. For his rule was not only absolute: it was also universal. "And the stars obey you?" "Certainly they do," the king said. "They obey instantly. I do not permit insubordination."
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, he plucked up his courage to ask the king a favour.

Source 3:
"What is this 'where I live,' of which you speak? Where do you want to take your sheep?"
After a reflective silence he answered: "The thing that is so good about the box you have given me is that at night he can use it as his house."
"That is so. And if you are good I will give you a string, too, so that you can tie him during the day, and a post to tie him to."
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."
Then the little prince said, earnestly: "That doesn't matter. Where I live, everything is so small."
And, with perhaps a hint of sadness, he added: "Straight ahead of him, nobody can go very far..."
I had thus learned a second fact of great importance: this was that the planet the little prince came from was scarcely any larger than a house! But that did not really surprise me much. I knew very well that in addition to the great planets, such as the Earth, Jupiter, Mars, Venus, to which we have given names, there are also hundreds of others, some of which are so small that one has a hard time seeing them through the telescope.

Observation: The Little Prince's relationship with his home planet is one of deep connection and care. He has a sense of responsibility towards it, which is reflected in his concern for the sheep that he left behind. The fact that he wants to watch the sunset multiple times without having to move his chair suggests that he values the comfort and familiarity of his own planet.
Moreover, when discussing his home planet, the Little Prince mentions that "Where I live, everything is so small!" This statement implies a sense of intimacy and closeness with his surroundings. He also seems to feel a pang of sadness when remembering his little planet, which he had forsaken, indicating a strong emotional bond with it.
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. This suggests that he takes his responsibility seriously and values the well-being of his planet and its inhabitants.
Overall, the Little Prince's relationship with his home planet is one of love, care, and responsibility, which is deeply ingrained in his sense of identity and purpose.

Thought: I can answer without using any more tools. I'll use the user's language to answer.

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. This relationship is deeply ingrained in his sense of identity and purpose.

Case 2. Using Llama 3.

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

Thought: The current language of the user is English. 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" (amor intellectualis Dei), and how does this concept function in his ethical system?"]

Source 1: But since the mind no longer endures, this pleasure does not involve affects and is not like the pleasures of the finite mind and body. 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" (P33C). This is not like love of another person, love of a thing or even the love of God of P15. Intellectual love of God is knowing and loving all being eternally (P33); it is the same love with which God loves itself and all

Source 2: 144 Spinoza's Ethics understands himself and his affects.' No idea, when it is truly understood, can be conceived without God (P15), so every adequate idea that we have is accompanied by the idea of God. 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 adequate ideas we have, the stronger is the idea of God and the stronger is our love for God. Our love for God grows with our true understanding. No one can hate God (P18), because our idea of God is necessarily adequate and cannot be accompanied by any kind of sadness. A person who believes he hates God is invariably thinking of an image of God (probably from religion) and does not understand God truly. Even when we understand that God is the cause of our sadness, we feel joy and love. For insofar as we understand God to be the cause of the sadness, we truly understand the sadness and it becomes joy (P18S). Our love for God is not met by God's love for us, for God experiences no affects (P17). God therefore is not jealous, envious or angry; he does not love or hate anyone. Our love of God is the most constant of the affects, for it is equivalent to true understanding itself and cannot be destroyed (P20S). Furthermore, the mind's greatest good is knowledge of God (IVP28), for true understanding is what is best for the mind. And so, because we desire others to gain true understanding, we desire that other people know and love God too (P20). Remembering that God is being, this "love of God" is a love of being. To love all being, and rejoice in it, is our greatest virtue.

Source 3: 154 Spinoza's Ethics its modes (P35, P36). Since God feels no affects, 'love' is perhaps not the best word; 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 (P36S).

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. Instead, it involves understanding and loving all being eternally. In this sense, the intellectual love of God is equivalent to true understanding itself and cannot be destroyed.
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. This love is not met by God's love for us, but rather, it is the affirmation of God's perfection and our own perfection in God, which constitutes our blessedness.
Ultimately, the intellectual love of God is considered the most constant of affects, and it is what is best for the mind. It is also seen as a love of being, where one rejoices in all being, making it our greatest virtue.

Thought: I can answer without using any more tools. I'll use the user's language to answer.

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.