

Methodology:

Content in form of essays written by Paul Graham is used. The essays are indexed to create vector embeddings which enable efficient retrieval of relevant content based on the queries made to the system. Streamlit(Front end framework) is used to create an interactive interface.

1. Content is indexed to create vector embeddings for semantic search
2. Semantic search is used to retrieve information from the content.
3. LLM used for natural language response.
4. (provides the answer which is now required to be tested)

Contents of Literature review:

The topics listed below would be expounded upon in the literature review.

1. Semantic search
2. Vector Embeddings
3. RAG
4. LLAMA
5. Mixture Of Experts (Model for
6. LLMs
7. Vector-Databases
8. Embeddings – help to search through the datasets
9. Inference and its definition
10. Evaluating Different LLMs
11. Justification behind choosing a certain architecture of embeddings model and LLM
12. Knowledge Graphs for Retrieval Augmented Models
13. RAG METRICS.

As it stands currently, I have researched the topics and have collected some resources. I will start formalizing the literature review after our meeting and getting a clearer picture of the objective and methodology.

Topics for discussion:

- Attribution for content which is being retrieved.
- Interference of world knowledge of LLM interfering with the content, can cause hallucinations.
- Would it be required to be staged and made available over the internet or are we just working on a proof of concept?
- Best ways to evaluate the responses? If we are looking at the semantic quality of answers in our applications then we can look into statistical scorers for evaluation such as BLEU(n-grams in LLM output tested against Ground Truth), ROUGE(overlap of n-grams) and other statistical scorers can be used.
- How to go about answering questions which fall outside the purview of what the system has been trained on.
- Not being able to work with open source AI models due to insufficient funds and API quota. OpenAI and Claude.

Below are snapshots from the current Prototype which uses Mistral as its LLM right now but has provision to use any OpenSource/Closed Source LLM.

Deploy

Chat with Paul Graham's Essays

Ask a question about Paul Graham's essay 'What I Worked On':

What is Pauls ideology on superlunar returns

Submit

User: What does Paul talk about when he speaks of the four quadrants of conformism?

Chatbot: In the given text, Paul discusses a way of classifying people based on their degree and aggressiveness of conformism using a Cartesian coordinate system. He labels the quadrants as follows: upper left - aggressively conventional-minded, lower left - passively conventional-minded, lower right - passively independent-minded, and upper right - aggressively independent-minded. The text describes characteristics and behaviors associated with each quadrant.

User: What is Pauls ideology on superlunar returns

Chatbot: Paul, as depicted in the text, appears to view superlinear returns as a phenomenon that arises from both curiosity-driven exploration and ambition, with curiosity potentially being more powerful. He also notes the dynamic nature of this territory and the potential for significant rewards from expanding it. Additionally, he emphasizes the importance of learning and knowledge, suggesting that they can lead to exponential growth and new discoveries that in turn yield superlinear returns. However, Paul cautions against blindly following popular or conventional paths, as there may be hidden reasons why these areas have mediocre returns despite high levels of competition. Overall, it seems that Paul sees superlinear returns as a result of pursuing interesting questions and making significant discoveries, rather than through traditional means such as climbing existing peaks or relying on competition alone.

Chat with Paul Graham's Essays

Ask a question about Paul Graham's essay 'What I Worked On':

what does Paul talk about schools?

response

Response

Response(response=' In the provided context, Paul talks about working on programming before college using an IBM 1401 computer, his experience with microcomputers and the TRS-80 that he eventually owned, and his attempts to study philosophy in college but switching to Artificial Intelligence (AI) du...

Response object.

Returned if streaming=False.

Attributes:

response: The response text.

metadata

dict

{'655e89fd-8cbc-4af6-a9c6-f5178a184cb1': {'file_path': 'data/paul_graham_essay.txt', 'file_name': 'paul_graham_essay.txt', 'file_type': 'text/plain', 'file_size': 75042, 'creation_date': '2024-02-16', 'last_modified_date': '2024-02-16', 'last_accessed_date': '2024-02-16'}, '2cf0da72-949a-445f-b01e-2...

response

str

' In the provided context, Paul talks about working on programming before college using an IBM 1401 computer, his experience with microcomputers and the TRS-80 that he eventually owned, and his attempts to study philosophy in college but switching to Artificial Intelligence (AI) due to his interests...

source_nodes

list

[NodeWithScore(node=TextNode(id_='655e89fd-8cbc-4af6-a9c6-f5178a184cb1', embedding=None, metadata={'file_path': 'data/paul_graham_essay.txt', 'file_name': 'paul_graham_essay.txt', 'file_type': 'text/plain', 'file_size': 75042, 'creation_date': '2024-02-16', 'last_modified_date': '2024-02-16', 'last...

get_formatted_sources

metho

Get formatted sources text.