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.

A screenshot of a chat

Description automatically generated

A screenshot of a computer

Description automatically generated