What is RAG:

RAG is an AI framework for retrieving facts from an external knowledge base to ground large language models (LLMs) on the most accurate, up-to-date information and to give users context relevant responses.

Applications of RAG:

1. RAG is a friendly way of interacting with large corpus of text data in time-efficient manner.
2. Querying research papers.
3. Exploring policy documents.
4. Drawing context relevant conclusions and insights.
5. Virtual Assistant (Chatbot).
6. Personalized recommendations.

Steps involved:

1. Document Parsing.
2. Database Creation.
3. Storing the embeddings.
4. Pipeline Creation.
5. Perform RAG.
6. Fine tune with DSPY.

Comparison of Offline (LLAMA) with online LLMs (GPT4)

|  |  |  |  |
| --- | --- | --- | --- |
| **SLNO** | **Feature** | **GPT** | **OLLAMA** |
| 1 | Versatility | Higher, with rich set of configurations available | Configurations are available only for large models after 13B parameters. |
| 2 | Memory | Online | A reasonably small model mistral with 5B parameters requires 7GB of exclusive RAM |
| 3 | Hallucinations | With proper prompting, hallucinations can be eliminated. | Hallucinations not under control. |
| 4 | Cost | Reasonable cost | Requires large computing power offline |
| 5 | Deployment on cloud | The only option | Not an option. The modest billing for AWS Ec2 instance is $850 per week. |
| 6 | Speed | Faster | slower |
| 7 | Data security | No data security | Absolute data security |
| 8 | Third Party Integrations | Langgraph and DsPY | Langgraph and DsPy |

**Cost Minimizing techniques**:

* Cost can be minimized by employing a range of hybrid techniques.
* Use Ollama text embeddings. Avoid GPT embeddings.
* For the GPT keep a limit on the number of queries per day.
* Keep a memory of the questions, when a similar question is asked (with a similarity above 90%), answer from the memory, instead using the LLM.
* Maintain a set of FAQs with corresponding LLM generated answers. Train the model on these FAQs. This way GPT costing will be only one time. After that provide the answers based on similarity of questions.