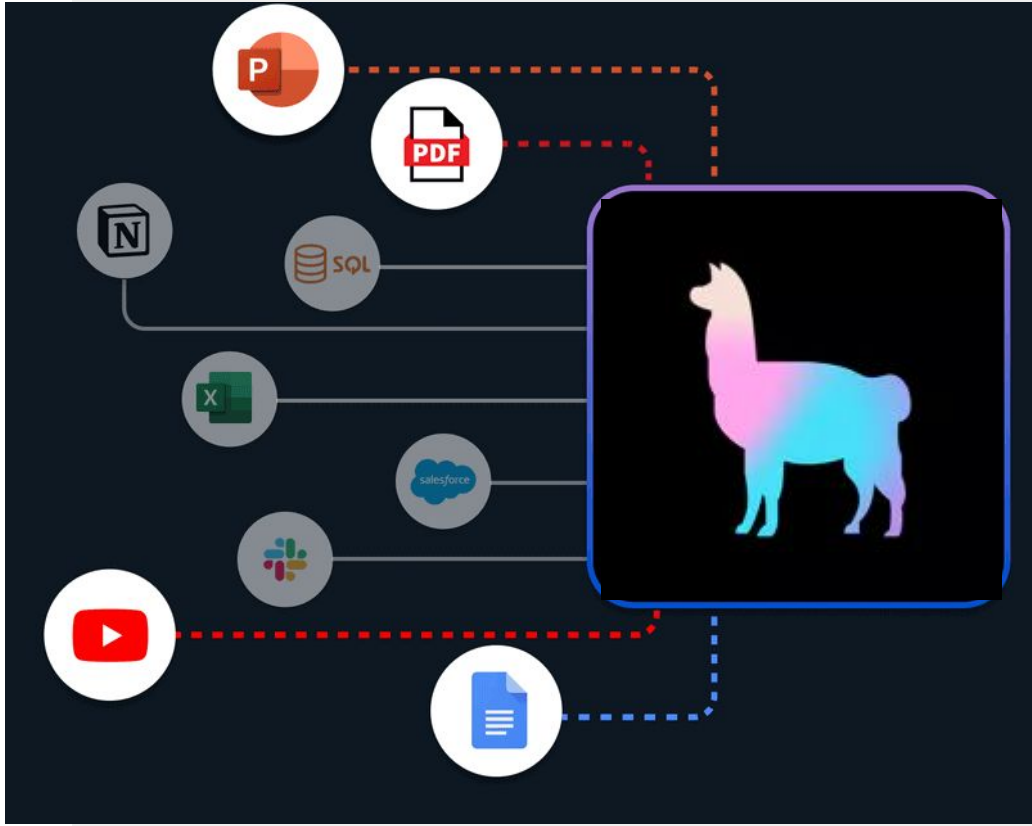
Create RAG apps for Enterprise Use Cases

Leveraging the power of advanced GPT models to develop apps tailored for enterprise needs.



Joydeep Bhattacharjee

Connect with ME!

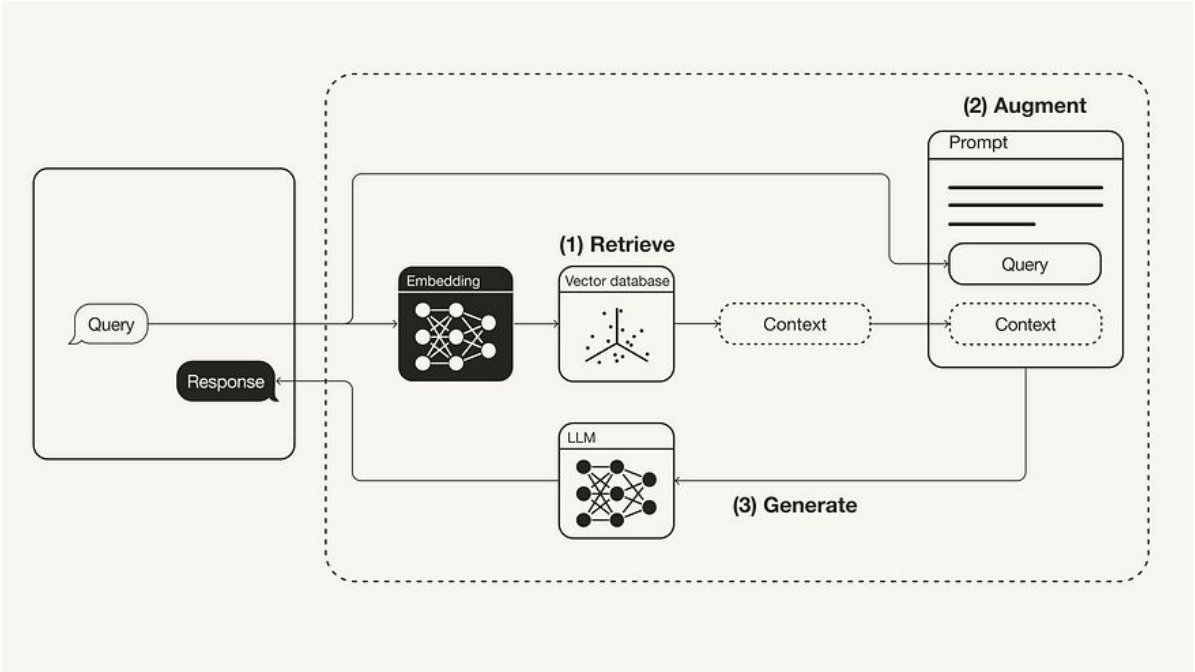
- NLP basics: https://vibrantai.academy/courses/1
- Career Guidance in Machine Learning: https://topmate.io/joydeep_bhattacharjee
- FREE Mock machine learning interview coach: https://vibrantai.academy/interview-trainer/chat
- Connect with me on linkedin to get updated on more and more FREE artifacts. https://www.linkedin.com/in/joydeep-bhattacharjee-934a1157/

Agenda

- What is RAG: retrieval augmented generation.
- Brief discussion on the different parts of the RAG pipeline.
- Showcase an example chatbot using streamlit, huggingface, llama-index and ollama.

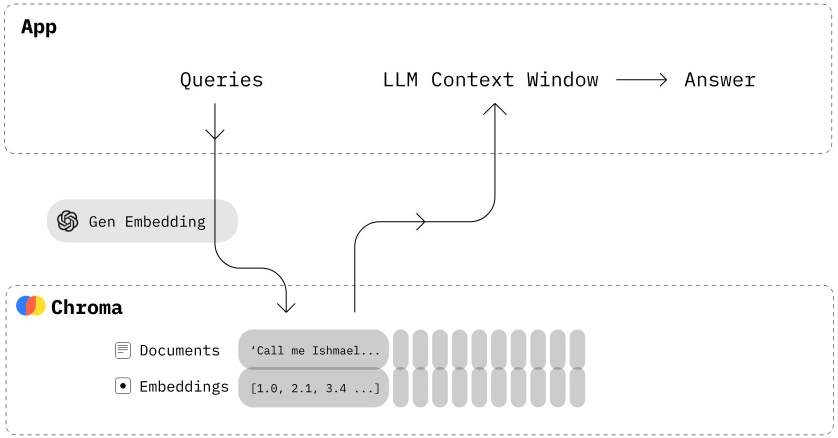
Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) is the concept to provide LLMs with additional information from an external knowledge source. This allows them to generate more accurate and contextual answers while reducing hallucinations.



Source: https://towardsdatascience.com/retrieval-augmented-generation-rag-from-theory-to-langchain-implementation-4e9bd5f6a4f2

Retrieval



Source: https://docs.trychroma.com/

- 3 parts: Generate embeddings + Vector DB + Reranking.
- Choose embeddings based on task whether retrieval or reranking. MTEB rankings.
- Lots of pretrained embeddings available. Commercial and Open Source
- Evaluation metrics: MRR and Hit rate.

Embedding Models

Model1: sentence-transformers/all-mpnet-base-v2

Model architecture and training: Mpnet base model taken and trained on 1B training pairs with contrastive learning objective.

Pooling: Mean pooling

Model2: SFR-Embedding-Mistral

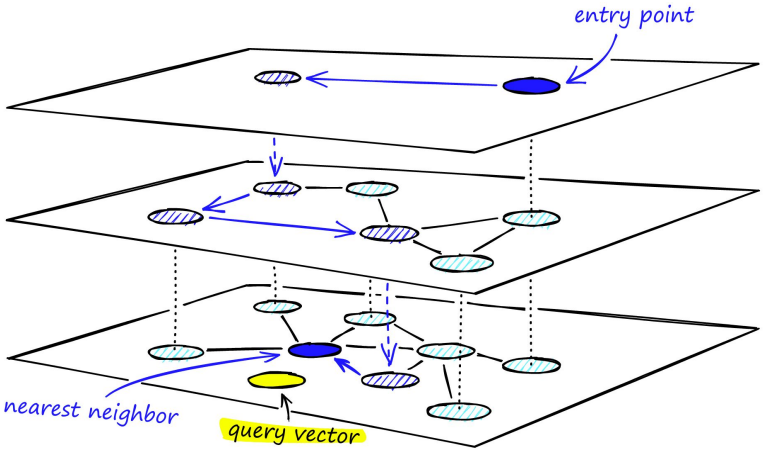
Pairs are not required for training. They used LLMs to generate synthetic data for a diverse range of text embedding tasks in 93 languages. Then use the task list and ask LLM to create query, pos, neg training data. Use this training data for training the embedding model.

Standard InfoNCE loss over the in-batch negatives and hard negatives.

Pooling: Last token pooling.

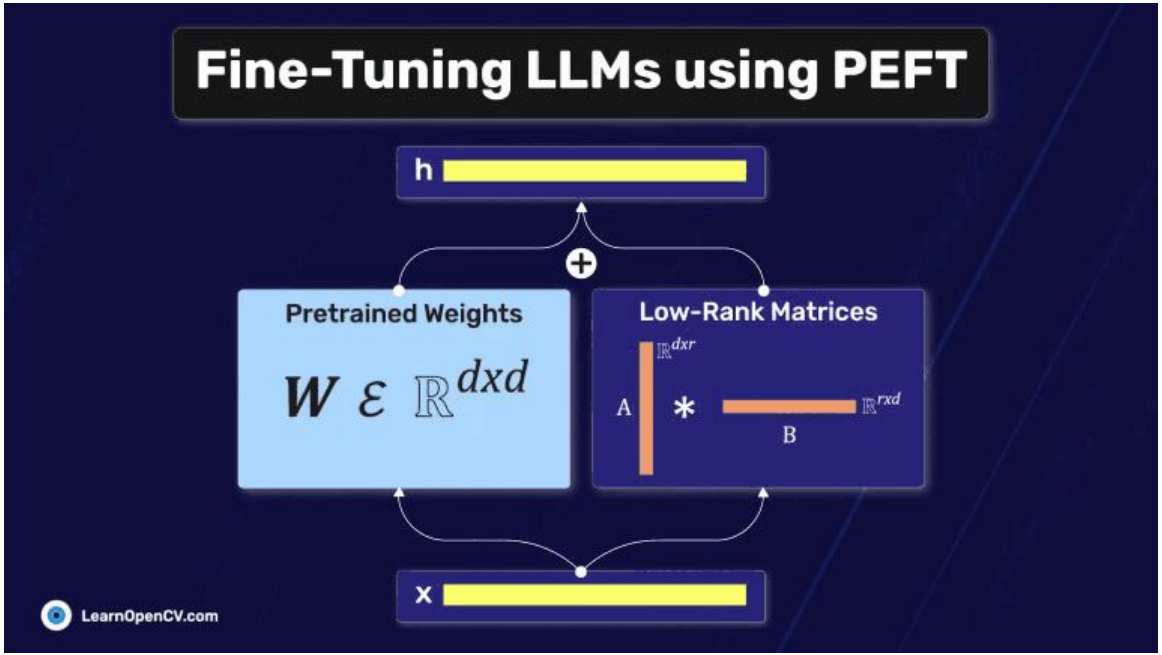
Vector DBs

- Many vector DBs in the market such as chroma, FAISS, pinecone etc. Older DBs also have vector tech.
- Consider technology aspects, enterprise readiness and developer experience when deciding vector DB.
- Popular algorithm HNSW but check the benchmarks. ANN-Benchmarks
- Benchmark on your own data.



Source: https://www.pinecone.io/learn/series/faiss/hnsw/

LLM Model Fine Tuning - LoRA (Low Rank Adaptation)



Source: https://learnopencv.com/fine-tuning-llms-using-peft/

Running models in production

- Try smaller parameter models. 2B vs 7B.
- Model Quantization.
- Create your own flask/faskAPI server.
- Ollama.
- Jan
- GPT4All

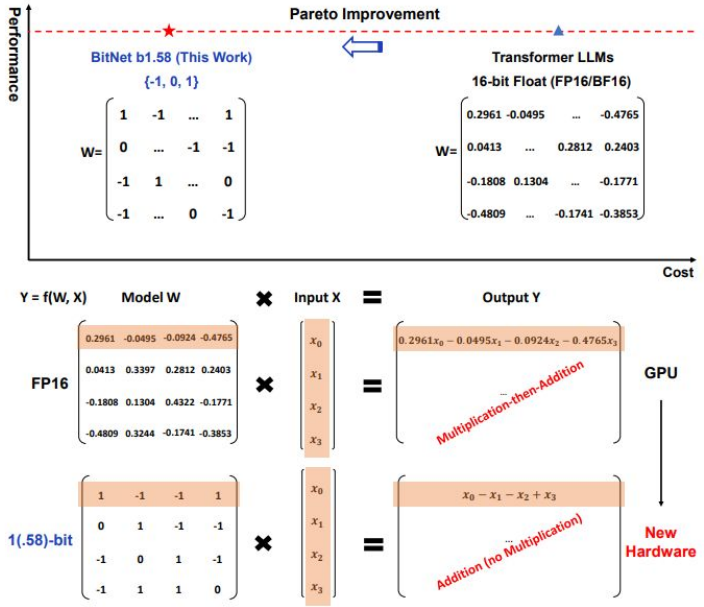


Figure 1: 1-bit LLMs (e.g., BitNet b1.58) provide a Pareto solution to reduce inference cost (latency, throughput, and energy) of LLMs while maintaining model performance. The new computation paradigm of BitNet b1.58 calls for actions to design new hardware optimized for 1-bit LLMs.

Source: https://arxiv.org/abs/2402.17764v1

RAG Evaluation

- Retrieval
  - Mean reciprocal rank
  - Hit rate
- LLM evaluation
  - Correctness: Correctness of a generated answer against a reference answer. For this you need the ground truth.
  - Faithfulness: Measure if the response from a query engine matches any source node.
  - Relevancy: Measure if the response + source nodes match the query.
- Observability

Code

https://github.com/infinite-Joy/webinars/tree/main/creating_gpt_chatbots_for_enterprise_usecases

References

- Code: https://github.com/infinite-Joy/webinars/tree/main/creating_gpt_chatbots_for_enterprise_usecases
- Llama-Index documentation: https://docs.llamaindex.ai/en/stable/getting_started/installation.html
- https://towardsdatascience.com/retrieval-augmented-generation-rag-from-theory-to-langchain-implementation-4e9bd5f6a4f2
- CM Insights: The large language model operations (LLMOps) market map - CB Insights Research
- LLM tools: https://github.com/underlines/awesome-ml/blob/master/llm-tools.md
- MTEB leaderboard: https://huggingface.co/spaces/mteb/leaderboard
- Boosting RAG: Picking the Best Embedding & Reranker models | by Ravi Theja | LlamaIndex Blog
- Approximate Nearest Neighbour Search Algorithm Benchmark: New approximate nearest neighbor benchmarks · Erik Bernhardsson
- An (Opinionated) Checklist to Choose a Vector Database | Pinecone