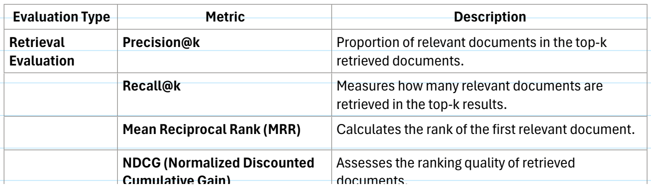
Capstone tasks - Groups

30 October 2024 07:33

Group1	Group2	Group3	Group4
Mohammed	Reeshabh	Hari	RohanJ
PInkesh	Anish	Tejas	Deepak
Pratyush	Aishwarya	RohanP	Sudeshna
Anup	Vidur	Vishal V	Komal
Sourav	Sripelly	Pruthvi	Navdeep
	Chiranjeevi		Eshwar
RAG (1) - analysis	Quantization based project (LoRA - just code explanations)	Evaluation of AI models (Transformers/ LM)	RAG (2) - Methods and algorithms
- COT/SC - Stuff/refine/map-reduce - similarity_search(query=query, k=5) - explain - Use OpenAI emb, plus 2 more emb models (HF) and show the top-k are similar or not - Evaluate the quality of the rtvr response (manual and LLM assisted one) - Simulate various queries and evaluate the answers (GLUE benchmarks..) Benchmark-GLUE-data...	- Quantization of VGGNet for Real-Time Image Classification - Reduce the memory footprint using quantization techniques. - Analyze the impact of quantization on accuracy, inference speed, and memory usage. - Explore mixed-precision quantization and layer-specific quantization strategies for further optimization. - Dataset and Preprocessing - Use a standard image classification dataset such as CIFAR-10, CIFAR-100 - Use a pretrained VGGNet (e.g., VGG16 or VGG19) from a framework like PyTorch or TensorFlow. - Evaluate the model's baseline performance on: ○ Accuracy. ○ Inference latency. ○ Memory usage. - Quantization Techniques a. Post-Training Quantization (PTQ) ○ Quantize the model to lower precision (e.g., INT8) after training. ○ Tools: TensorFlow Model Optimization Toolkit, PyTorch torch.quantization. b. Quantization-Aware Training (QAT): - Fine-tune the model while simulating quantization effects during training. - Tools: TensorFlow tf.quantization or PyTorch QuantizationAwareTraining. - Compare the quantized models against the baseline using: - Accuracy: Top-1 and Top-5 accuracy on the test dataset. - Memory Usage: Model size reduction.	- Standard metrics (Perplexity, BLEU, ROUGE, METEOR) - Benchmarking framework 1. GLUE 2. SuperGLUE 3. HELM ** 4. Big-bench **	- Stuff/refine/map-reduce - Libraries are there (Langchain, LlamaIndex) - Techniques for faster embedding creation - Batch - async - Efficient Vector Stores - Indexing techniques - Advanced Retrieval Techniques - Dual encoders - Cross encoders - Addressing Diversity: Maximum Marginal Relevance (MMR) - Embedding Adapters - Response Synthesis Optimization - Fine-tuning (illustrations) - Quantization - PEFT Benchmark-GLUE-data... OR any other PDF of your choice

RAG
- Understanding the PDF/CSV
  - Design queries
- Loading
  - Reader
  - Nodes and documents
- Indexing
  - Indexing choices



Evaluation Type	Metric	Description
Retrieval Evaluation	Precision@k	Proportion of relevant documents in the top-k retrieved documents.
	Recall@k	Measures how many relevant documents are retrieved in the top-k results.
	Mean Reciprocal Rank (MRR)	Calculates the rank of the first relevant document.
	NDCG (Normalized Discounted Cumulative Gain)	Assesses the ranking quality of retrieved documents.

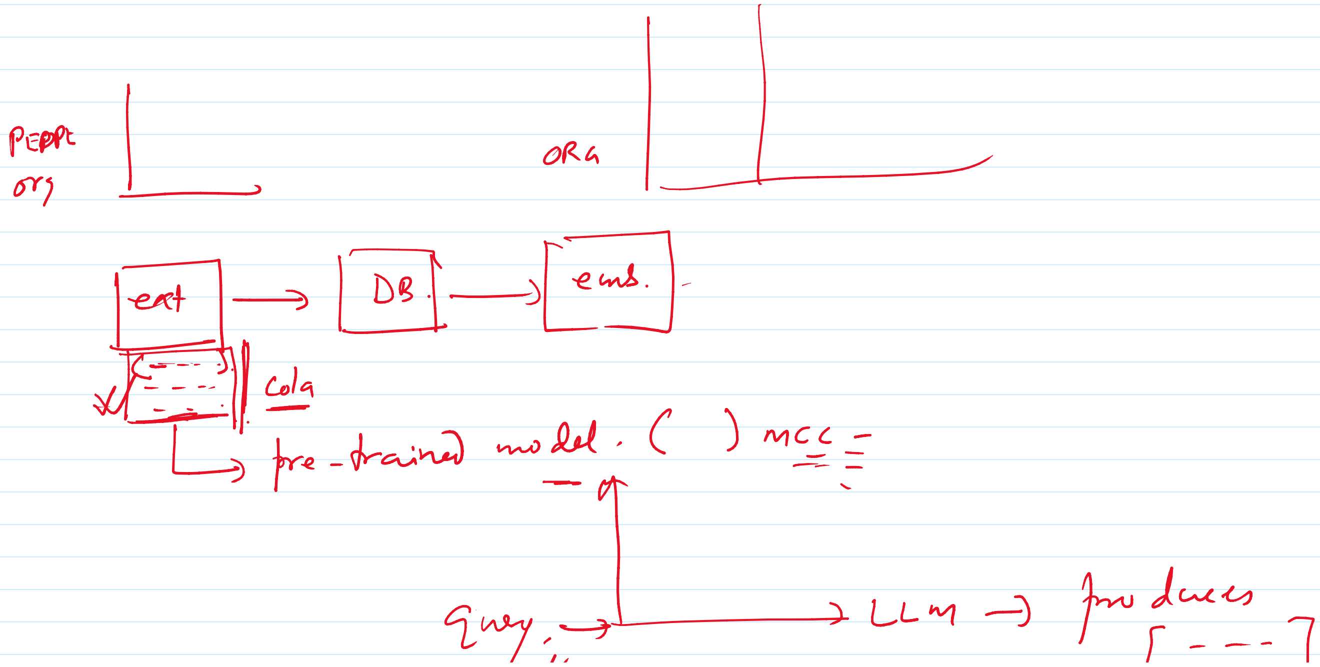
Reader
Nodes and documents
Indexing
Indexing choices
Embeddings
Storing
Vector Stores
Retrieval algorithms
Query
Retrievers
Routers
Post processors
Response synthesizers
Evaluation

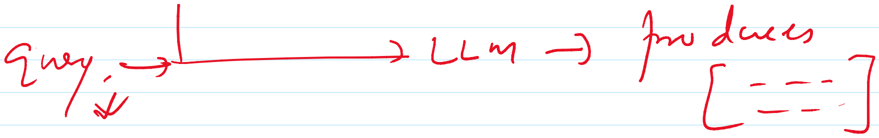
Applications
Query Engines/ Chat Engines
Agents

	Recall@k	Measures how many relevant documents are retrieved in the top-k results.
	Mean Reciprocal Rank (MRR)	Calculates the rank of the first relevant document.
	NDCG (Normalized Discounted Cumulative Gain)	Assesses the ranking quality of retrieved documents.
	MAP (Mean Average Precision)	Average precision for all queries, capturing overall retrieval quality.
Generation Evaluation	**BLEU**	Measures precision of n-grams between generated and reference texts.
	ROUGE	Measures recall of n-grams in the generated text compared to reference texts.
	METEOR	Combines precision and recall with additional features like synonym matching.
	Perplexity	Measures how well the model predicts the next token in a sequence.
End-to-End Evaluation	**Human Evaluation**	Subjective evaluation on fluency, coherence, informativeness, relevance, etc.
	Diversity and Novelty	Measures how diverse and novel the generated responses are.

RAG and **Advanced** RAG
Transformer and Fine tuning (Quant)
RLHF (basics of RL - Policy Gradient algo, PPO)
Image Generation
KG based retrieval
LangGraph (Agentic workflow)
Chatbot

LLM parameters (OpenAI)




Query
LLM
produces