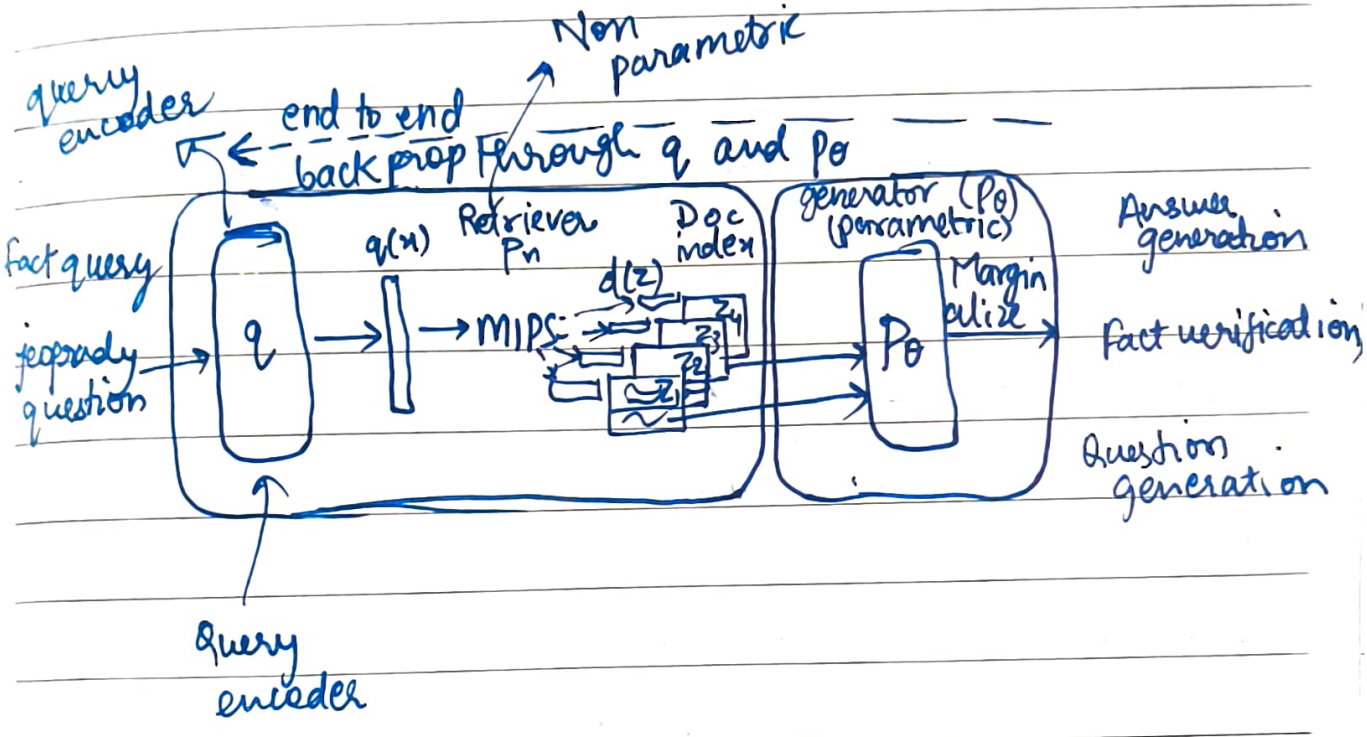
"Retrieval-Augmented Generation for knowledge-Intensive NLP Tasks"



→ combines a parametric pre trained model & a non parametric retriever.

Architecture ★

⇒ Input Query → denoted as 'x'. Can be a question, a fact to verify or an entity.

⇒ Query encoder → inputs 'x' → encoded in a dense vector representation generally ~~encoded~~ based on a BERT

→ Retriever (non parametric)

↳ uses Maximum Inner Product Search (MIPS) to find top K matching documents, documents which have highest dot product with the query vector indicating relevance.

→ Here retriever uses a DPR (Dense Passage Retriever) and bi-encoder architecture (BERT)

→ Document Index → non parametric memory, storing embeddings of pre-processed & indexed documents.

→ Generator → (Parametric memory)

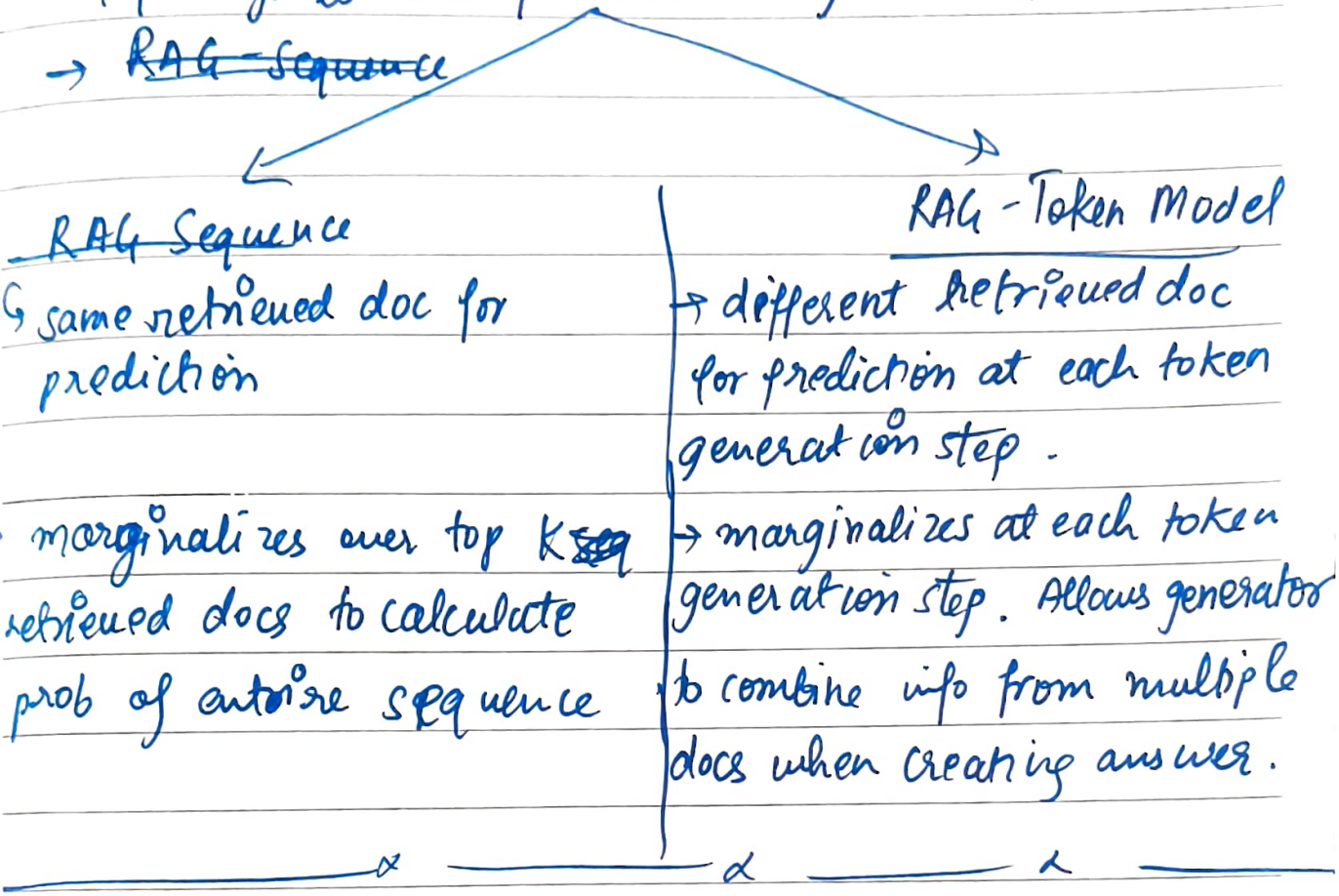
→ seq 2 seq model, takes 'x' & retrieved docs 'z'. pre traind BART model used here Generates output response denoted as 'y'

→ Output Generation → retrieved docs 'z' are treated as latent variable, and model marginalize over these docs to produce the probability distribution for generated text.

↳ the one not directly observed or explicitly chosen as a part of output but rather its possible values are considered to determine the final outcome.

Paper gives two formulations of the RAG model

→ ~~RAG-Sequence~~



MIPS → method for finding top K most matching relevant docs.
→ finds highest dot product, in sub linear time

Parametric memory → stored within models trainable parameters. (θ). Models that rely on parametric memory are limited by the knowledge in their training data upto to its cutoff point. lesser hallucinations

Non parametric memory → external knowledge, explicit, retrieval based memory. Knowledge can be easily updated (DPR)