Mathematical Representation of RAG

RAG is modeled as a combination of **retrieval** and **generation** processes. It retrieves relevant documents from an external knowledge base and generates responses conditioned on both the input query and retrieved documents.

Step 1: Retrieval of Relevant Documents

Given an input query q, the goal of retrieval is to find a set of documents $D = \{d_1, d_2, \ldots, d_n\}$ from a knowledge base K that maximizes relevance:

$$P(D|q) = \prod_{i=1}^{n} P(d_i|q)$$

Step 2: Conditional Response Generation

The generation process is modeled to produce a response r conditioned on both the input query q and the retrieved documents D:

$$P(r|q, D) = \sum_{d \in D} P(r|q, d)P(d|q)$$

Here: - $P(r|q, D)$ is the probability of generating the response given the query and documents. - $P(d|q)$ is the probability of selecting document d given the query. - $P(r|q, d)$ is the probability of generating the response conditioned on a specific document d and the query.

Combined Objective for RAG

The RAG framework combines both steps as:

$$P(r|q) = \sum_{d \in D} P(r|q, d)P(d|q)$$

Thus, the final response is generated by marginalizing over all retrieved documents.