https://arxiv.org/pdf/2005.11401.pdf

Combine pre trained parametric and nonparametric memory for language generation

- Parametric memory is a pretrained seq2seq model
- Non parametric memory is a dense vector index of Wikipedia, accessed with a pretrained neural retriever
- General purpose fine tuning approach (RAG)

Pretrained LLMS have been shown to store factual knowledge in their parameters and achieve state of the art results when fine tuned on downstream NLP tasks

- Ability to access and precisely manipulate knowledge is still limited and hence on knowledge intensive tasks their performance lags behind task specific architectures
- Finding place of origin for their decisions and updating their world knowledge remain open research problems

Pretrained neural language models learn a substantial amount of indepth knowledge from data

- Do so without any access to an external memory as a parameterized implicit knowledge base
- Cannot expand or revise their memory, can't provide insights into their predictions and may "hallucinate"

Hybrid models (parametric memory with non parametric [retrieval based] memory) can address these issues because knowledge can be directly revised and expanded and accessed knowledge can be inspected and interpreted

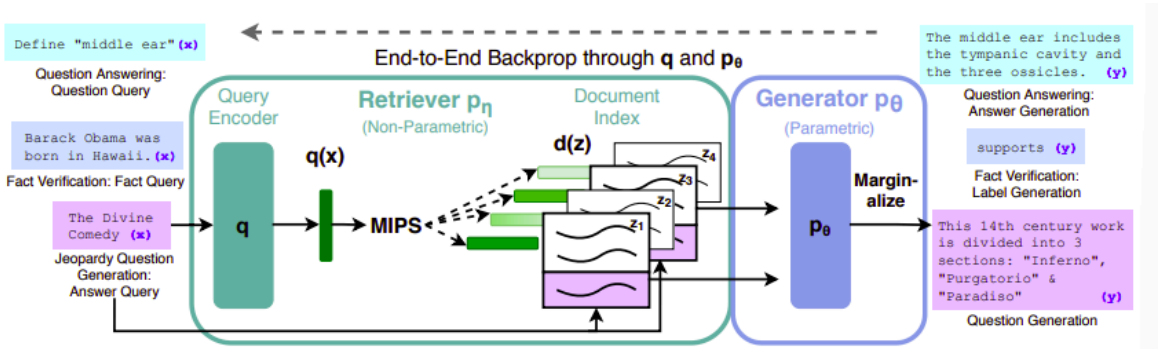

Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder + Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

Dense Passage Retriever (DPR): provides latent documents conditioned on the input
Generator/ parametric (Seq2Seq [BART]) conditions on these latent documents together with the input to generate the output
Non parametric: dense vector index of Wikipedia accessed with retriever
Marginalize latent documents with top- K approximation either per output (same document responsible for all tokens) or per token (different documents are responsible for different tokens)

By using pre trained access mechanisms, the ability to access knowledge is present without additional training

Knowledge intensive tasks- tasks that humans could not reasonably be expected to perform without access to an external knowledge source

Treat retrieved document as a latent variable
Generator: $p\theta(yi\ |x, z, y1{:}i{-}1)$
Retriever: $p\eta(z|x)$

Calculating top-k, the list of k documents z with highest prior probability is a Maximum Inner Product Search