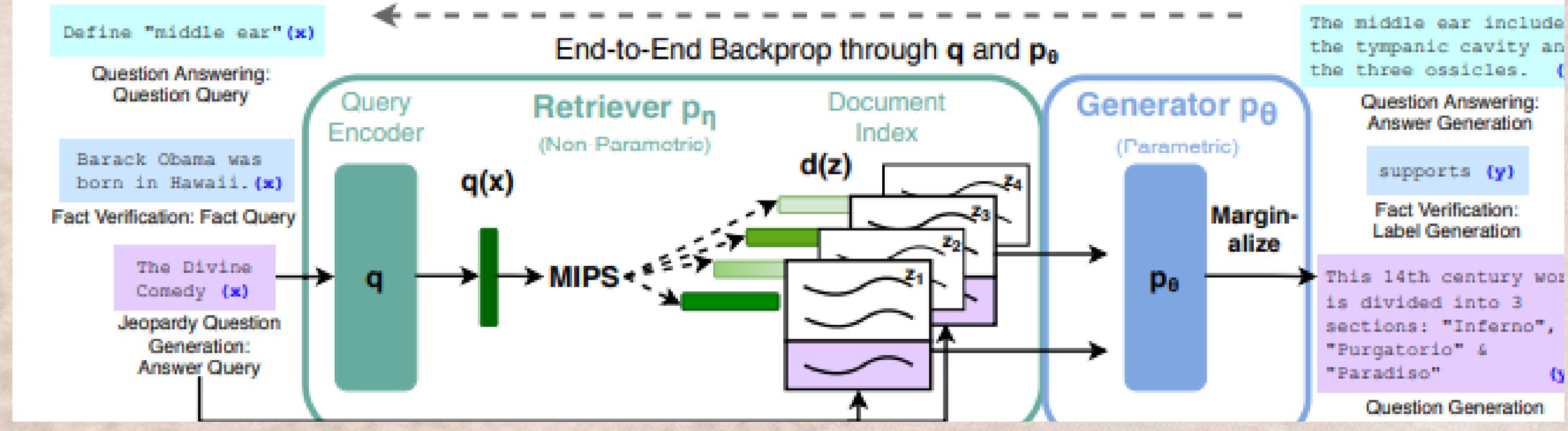
sound emotion embedding

EMOTION BASE RAG



MAIN IDEA

We utilize BERT base models for emotion tagging, processing an embedded sound token alongside an ASR model tailored to the specific domain. For each token, we get an emotion-embedded latent variable. This variable is then concatenated with the remaining token variables to form a sequence.

This process is analogous to the retriever component in a RAG model, which creates a document embedding for each token and appends it to the sequence. However, in our case, the appended latent variable represents emotional information instead of a retrieved document. Essentially, at each state, we incorporate two distinct latent variables. The remaining model details, such as the training methodology, remain consistent with the standard RAG framework.

RETRIEVAL AUGMENTED GENERATION

RAG models utilize an input sequence xx to retrieve text documents zz, which are then used as additional context for generating the target sequence yy. The architecture consists of two main components: (i) a retriever $p_\eta(z|x)$ $p_\eta(z|x)$, parameterized by ηη, which returns a top-K truncated distribution over text passages given the query xx; and (ii) a generator $p_\theta(y_i|x,z,y_{1:i-1})$ $p_\theta(y_i|x,z,y_{1:i-1})$, parameterized by θθ, which generates the current token based on the original input xx, the retrieved passage zz, and the context of the previously generated tokens $y_{1:i-1}$ $y_{1:i-1}$. To enable end-to-end training of both the retriever and the generator, the retrieved document zz is treated as a latent variable. The retrieval component $p_\eta(z|x)$ $p_\eta(z|x)$ is based on DPR (Dense Passage Retrieval), which uses a biencoder architecture built on BERT.