1. Data Preparation: We'll process the US section of the MedQA dataset and split the textbooks into manageable sections.
2. Embedding and Indexing: We'll use a pretrained embedding model to create vector representations of questions, answers, and textbook sections, then store them in a vector database.
3. Hybrid Search: We'll implement both dense and sparse retrieval methods, then combine them using a variable alpha parameter for flexibility.
4. Question Answering Model: We'll use a pretrained LLM to generate answers based on the retrieved context.
5. Streamlit UI: A basic chatbot interface using Streamlit and Langchain for easy interaction with the system.
6. Evaluation: We'll use the provided test set to assess the system's performance.

Finding the best combination of data retrieval from textbooks for our question context

1. Pointwise Mutual Information (PMI)

Definition: PMI is a statistical measure that quantifies the association between two n-grams (sequences of n items from a given sample of text) by comparing their joint probability to the product of their individual probabilities. The formula for PMI is:

$$PMI(x,y)=\log(p(x,y)/\ p(x)p(y)$$

- n-grams: These are contiguous sequences of n items from a given text or speech. For example, unigrams are single words, bigrams are pairs of consecutive words, and trigrams are triplets.
- p(x): The probability of n-gram x occurring in the document collection C.
- p(y): The probability of n-gram
- y occurring in C.
- p(x, y): The joint probability that both n-grams x and y occur together within a specified window (in this case, a 10-word window).

2. Info Retrieval

Information Retrieval involves obtaining relevant information from a large repository based on user queries. In this context, it uses a standard text retrieval system built on Apache Lucene or Elasticsearch.

Key Components:

- Inverted Index: A data structure that maps content (like words or terms) to its locations in a database file or document. This allows for fast full-text searches.

- BM25 Ranking: A probabilistic model used to rank documents based on their relevance to a search query. It considers term frequency and document length among other factors.

3 .**MaxOUT Mode**l-

Uses BiGRU to encode context and questions and does Max pooling to create final representation vectors.

$$p(q, a_i|c) = W_1(\tanh(W_2 h)) \in \mathbb{R}^1$$