***Text Splitting Parameters***

*text\_splitter = RecursiveCharacterTextSplitter(*

*chunk\_size=500, # Size of each text chunk*

*chunk\_overlap=50, # How much overlap between chunks*

*add\_start\_index=True, # Whether to add start index metadata*

*separators=["\n\n", "\n", " ", ""] # How to split the text*

*)*

***Vector Store Retrieval Parameters***

*retriever = vector\_store.as\_retriever(*

*search\_type="similarity", # Type of search ("similarity" or "mmr")*

*search\_kwargs={*

*"k": 4, # Number of documents to retrieve*

*"include\_metadata": True,*

*"score\_threshold": 0.5 # Minimum similarity score threshold*

*}*

*)*

***LLM Parameters***

*llm = ChatOpenAI(*

*temperature=0 # Controls randomness (0 = deterministic, 1 = more random)*

*# model="gpt-4" # You can specify different models*

*# max\_tokens=None # Maximum tokens in response*

*# top\_p=1.0 # Nucleus sampling parameter*

*# frequency\_penalty=0 # Reduces repetition of tokens*

*# presence\_penalty=0 # Encourages new topics*

*)*

***RetrievalQA Chain Parameters***

*qa\_chain = RetrievalQA.from\_chain\_type(*

*llm=llm,*

*chain\_type="stuff", # Can be "stuff", "map\_reduce", "refine", or "map\_rerank"*

*retriever=retriever,*

*return\_source\_documents=True,*

*verbose=True*

*)*

***Key parameters and their effects:***

***1. Text Splitting:***

*chunk\_size: Larger chunks provide more context but may reduce relevance*

*chunk\_overlap: More overlap helps maintain context across chunks*

*separators: Controls how text is split into chunks*

***Retrieval:***

*search\_type: "similarity" or "mmr" (Maximum Marginal Relevance)*

*k: Number of relevant documents to retrieve*

*score\_threshold: Minimum similarity score for retrieved documents*

***LLM:***

*temperature: Higher values (0-1) make output more creative/random*

*model: Different OpenAI models (e.g., gpt-3.5-turbo, gpt-4)*

*max\_tokens: Limit response length*

*frequency\_penalty: Reduce repetition*

*presence\_penalty: Encourage topic diversity*

***Chain:***

*chain\_type: Different methods for processing retrieved documents*

*verbose: Toggle detailed processing information*

*To improve results, you might:*

*Increase chunk\_size for more context*

*Increase k to retrieve more documents*

*Adjust score\_threshold to control relevance*

*Try different chain\_type options*

*Experiment with temperature for different response styles*