RAG Framework Evaluation

### 1. Chunking (Splitting Documents)

Chunking in a RAG involves breaking down large documents into smaller chunks for efficient retrieval and better model input handling.

* **Method 1: Fixed-Length Chunking**
  + Pros: Simple to implement, good for maintaining context.
  + Cons: May break semantic coherence if chunks are not aligned with sentence or paragraph boundaries.
* **Method 2: Recursive Character Text Splitter**
  + Pros: Preserves semantic coherence by splitting at sentence or paragraph boundaries; prevents breaking of meaning.
  + Cons: More computationally expensive than fixed-length chunking.
* **Method 3: Custom Delimiters (e.g., split by headings)**
  + Pros: Chunks are more contextually meaningful, ideal for structured documents like manuals.
  + Cons: Not applicable for less structured or conversational documents.

### 2. Embedding Methods (Document Representation)

Embeddings represent chunks as vectors that models can process and compare.

* **Method 1: Sentence Transformers (e.g., all-MiniLM-L6-v2)**
  + Pros: Efficient and fast for many use cases, optimized for sentence-level similarity.
  + Cons: May lose performance with very long texts or complex semantic relationships.
* **Method 2: OpenAI Embeddings (e.g., GPT-3 based)**
  + Pros: High-quality embeddings due to large model size, great for complex contexts.
  + Cons: High latency, costs associated with API usage.
* **Method 3: BERT-Based Embeddings**
  + Pros: Good balance between performance and efficiency, useful for general-purpose tasks.
  + Cons: Often requires fine-tuning for task-specific performance.

### 3. Vector Databases (for storing embeddings)

* **Method 1: Chroma**
  + Pros: Easy to use, specifically optimized for document-based tasks, integrated well with LangChain.
  + Cons: Limited scalability compared to enterprise solutions.
* **Method 2: FAISS**
  + Pros: Highly scalable, customizable, optimized for fast similarity search on large datasets.
  + Cons: Requires more setup, lacks advanced features like metadata support.
* **Method 3: Pinecone**
  + Pros: Managed service, scalable without setup, supports metadata for filtering results.
  + Cons: Commercial service, can incur costs.

### 4. Document Retriever Methods (Retrieving relevant chunks)

* **Method 1: Similarity-Based Retrieval**
  + Pros: Fast and simple, great for tasks with well-defined queries.
  + Cons: May retrieve irrelevant documents if the query is vague.
* **Method 2: BM25 (Lexical Search)**
  + Pros: Strong for keyword matching, handles large corpora well.
  + Cons: Cannot capture semantic similarity; weaker on nuanced queries.
* **Method 3: Hybrid Retrieval (BM25 + Embeddings)**
  + Pros: Combines the strengths of both semantic and keyword search.
  + Cons: More computational overhead due to dual search modes.

### 5. Similarity Search Methods (Comparing query with chunks)

* **Method 1: Cosine Similarity**
  + Pros: Fast and works well for text embeddings.
  + Cons: Not suitable for cases where magnitude of vectors matters.
* **Method 2: Euclidean Distance**
  + Pros: Simple and interpretable.
  + Cons: Less effective when vectors are high-dimensional, doesn't normalize vectors.
* **Method 3: Dot Product Similarity**
  + Pros: Efficient for large-scale similarity tasks, commonly used in models like BERT.
  + Cons: Sensitive to vector length, so may need normalization.

### Table for Comparison

| **Aspect** | **Method** | **Pros** | **Cons** |
| --- | --- | --- | --- |
| **Chunking** | Fixed-Length | Simple, maintains some context | May break meaning across boundaries |
|  | Recursive Character Text Splitter | Preserves meaning, more granular control | More computationally expensive |
|  | Custom Delimiters | Contextually meaningful chunks | Limited applicability to structured docs |
| **Embedding** | Sentence Transformers | Efficient, good for sentence-level tasks | Struggles with complex semantics |
|  | OpenAI Embeddings | High-quality, good for complex contexts | Latency, costs |
|  | BERT-Based Embeddings | Balance of efficiency and performance | Needs fine-tuning |
| **Vector Databases** | Chroma | Easy-to-use, document-focused | Scalability limitations |
|  | FAISS | Highly scalable, customizable | Requires more setup |
|  | Pinecone | Managed service, metadata support | Costs |
| **Document Retriever** | Similarity-Based Retrieval | Fast, effective for precise queries | Can struggle with vague queries |
|  | BM25 | Good for keyword matching | No semantic similarity |
|  | Hybrid (BM25 + Embeddings) | Best of both worlds | More computationally expensive |
| **Similarity Search** | Cosine Similarity | Fast, works well for embeddings | Magnitude doesn't matter, limited for certain cases |
|  | Euclidean Distance | Simple, interpretable | Not good for high-dimensional vectors |
|  | Dot Product Similarity | Efficient for large-scale tasks | Sensitive to vector length, needs normalization |

### Next Steps

It is recommended to experiment with these methods in RAG projects to see which combinations perform best based on the specific use case. Each method has its strengths and trade-offs, so the ideal choice depends on factors like dataset size, query complexity, and performance requirements.