**Assessment Task: Question-Answering using Open-Source GDPR Dataset**

**Overview**

This document outlines assessment task focused on building a Retrieval-Augmented Generation (RAG) pipeline using the General Data Protection Regulation (GDPR) dataset from the EUR-Lex database. The following is the expected expertise:

* Embeddings & Similarity Search (Cosine Similarity, BM25, Hybrid Search, Vector Search)
* Chunk Size Optimization & Addressing "Lost in the Middle" Problem
* Summarization & Follow-Up Handling
* End-to-End Question-Answering (QA) Pipeline Implementation

Candidate will work on retrieving insights based on different types of queries, optimizing retrieval accuracy, and summarizing text effectively.

**Categories of Queries**

The assessment includes four types of legal queries:

1. Fact-Based Queries – Direct retrieval of explicit information from GDPR.
2. Abstract Queries – Concept-based queries that require a deeper understanding.
3. Reasoning-Based Queries – Queries that require inferences across multiple GDPR sections.
4. Comparative Queries – Cross-section analysis within GDPR or comparisons with other laws.

Candidates must ensure that their system efficiently retrieves relevant sections, ranks them correctly, and presents responses in a structured format.

**Queries for Assessment**

| Category | Query |
| --- | --- |
| Fact-Based | 1. What is the maximum fine for a GDPR violation? |
|  | 2. What does GDPR say about the right to be forgotten? |
|  | 3. Under what conditions can personal data be transferred outside the EU? |
|  | 4. What is the role of a Data Protection Officer (DPO)? |
|  | 5. What constitutes a ‘data breach’ under GDPR? |
|  | 6. What are the lawful bases for processing personal data under GDPR? |
|  | 7. How long can organizations retain personal data under GDPR? |
|  | 8. What are the key rights of individuals regarding their personal data? |
|  | 9. What constitutes valid consent under GDPR? |
|  | 10. What obligations do data processors have under GDPR? |
| Abstract | 11. How does GDPR define data minimization? |
|  | 12. Why is GDPR considered a landmark regulation in data privacy? |
|  | 13. What are the key differences between explicit and implicit consent under GDPR? |
|  | 14. How does GDPR affect AI-based data processing? |
| Reasoning-Based | 15. If a company stores customer data without informing users, which GDPR articles does it violate? |
|  | 16. Can a company use personal data without consent if they anonymize it? |
|  | 17. If a user deletes their account, does GDPR require their data to be erased immediately? |
| Comparative | 18. How does GDPR differ from the California Consumer Privacy Act (CCPA)? |
|  | 19. How does GDPR handle children’s data protection compared to COPPA? |
|  | 20. What are the key similarities between GDPR and Brazil’s LGPD? |

**Techniques to Explore**

Candidates are encouraged to explore the following techniques (or choose technique of your choice) for optimizing retrieval and summarization:

1. Embedding & Similarity Techniques

* Cosine Similarity: Measures angular distance between document embeddings.
* Hybrid Search (BM25 + Embeddings): Combines lexical and semantic search for improved accuracy.
* Vector Search Techniques:
  + FAISS (Facebook AI Similarity Search): Efficient nearest-neighbor search for high-dimensional vectors.

2. Chunk Size Optimization

* Sliding Window Approach: Ensures contextual information is preserved.
* Optimal Token Length Testing: Helps address the “Lost in the Middle” problem in LLMs.

3. Summarization Models

* Abstractive Summarization (GPT, Llama): Generates summaries in natural language.

4. Handling Multi-Turn Queries

* Memory-Augmented Retrieval: Maintains query history for better context.
* Conversational Agents (LangChain, RAG Pipelines): Enables follow-up question handling using conversational buffer memory.

**Expected Deliverables**

* Python Codebase: Implementation using LangChain/OpenAI API.
* Notebook Analysis: Explanation of similarity techniques and evaluation.
* Performance Comparison: Results from multiple retrieval methods.

This assessment will provide insights into a candidate’s ability to work with NLP, embeddings, RAG pipelines, and optimizing information retrieval.