Middle East Technical University
Informatics Institute

# ASSESSING ADVANCED RETRIEVAL-AUGMENTED GENERATİON TECHNİQUES FOR QUESTION ANSWERING: A CASE STUDY ON GOVERNMENTAL SERVICES

Advisor Name: Prof. Dr.Tuğba Taşkaya Temizel
(METU)

Student Name: Abdullah Talat Ahmed Alzariqi
(Information Systems)

January 2025

TECHNICAL REPORT
METU/II-TR-2025-

Orta Doğu Teknik Üniversitesi
Enformatik Enstitüsü

# SORU-CEVAP İÇİN GELİŞMİŞ ALMA-ARTTIRMALI ÜRETİM TEKNİKLERİNİN DEĞERLENDİRİLMESİ: DEVLET HİZMETLERİ ÜZERİNE BİR VAKA ÇALIŞMASI

Danışman Adı: Prof. Dr.Tuğba Taşkaya Temizel
(ODTÜ)

Öğrenci Adı: Abdullah Talat Ahmed Alzariqi
(Bilişim Sistemleri)

Ocak 2025

TEKNİK RAPOR
ODTÜ/II-TR-2025-

# REPORT DOCUMENTATION PAGE

| | |
|---|---|
| **1. AGENCY USE ONLY (Internal Use)** | **2. REPORT DATE**<br>16/01/2025 |

**3. TITLE AND SUBTITLE**

**ASSESSING ADVANCED RETRIEVAL-AUGMENTED GENERATİON TECHNİQUES FOR QUESTION ANSWERING: A CASE STUDY ON GOVERNMENTAL SERVICES**

| | |
|---|---|
| **4. AUTHOR (S)**<br>Abdullah Talat Ahmed Alzariqi | **5. REPORT NUMBER (Internal Use)**<br>**METU/II-TR-2025-** |

**6. SPONSORING/ MONITORING AGENCY NAME(S) AND SIGNATURE(S)**
Informatics Non-Thesis Master's Programme, Department of Information Systems, Informatics Institute, METU

Advisor: Prof. Dr.Tuğba Taşkaya Temizel Signature:

**7. SUPPLEMENTARY NOTES**

**8. ABSTRACT (MAXIMUM 200 WORDS)**

This case study evaluates advanced Retrieval-Augmented Generation (RAG) methods to enhance question answering in conversational AI for government services, focusing on data from the Ministry of Health and Prevention. It assesses various RAG strategies like naive RAG, HyDE, Hybrid RAG, Corrective RAG, Self RAG, and Astute RAG to tackle issues like hallucinations and reasoning gaps. The study's methodology includes three phases: benchmarking embedding models, testing LLMs, and evaluating RAG pipelines. Results show Astute RAG excels in improving accuracy and context alignment, providing insights for optimizing AI solutions in government and healthcare, while highlighting limitations in dataset and evaluation methods.

| | |
|---|---|
| **9. SUBJECT TERMS**<br>**RAG, Evaluation, RAG Triad,** | **10. NUMBER OF PAGES**<br>**68** |

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

## Introduction:

Retrieval-Augmented Generation (RAG) is an approach that combines the power of information retrieval and generative language models. RAG models integrate two key components: a retriever and a generator. The retriever searches a knowledge base for the most relevant documents or passages in response to a query, while the generator processes the retrieved information to generate more accurate and contextually relevant responses. This hybrid system leverages the advantages of retrieval-based systems, such as precise information retrieval, and generative models, which can provide human-like language generation capabilities (Lewis et al., 2020). RAG is particularly useful for scenarios where the generative model lacks up-to-date or domain-specific knowledge, as the retriever can provide it with relevant context.

RAG has been widely applied in various domains, including customer support, conversational AI, code generation, and knowledge-based question answering (QA). For instance, in customer support, RAG can help generate responses based on previously answered tickets and existing knowledge bases, ensuring accuracy and consistency. In legal and medical fields, RAG can assist professionals by retrieving case-specific information or medical literature, and then generating tailored summaries (Izacard & Grave, 2021). Additionally, RAG's ability to work with structured databases makes it suitable for generating SQL queries or other structured outputs based on natural language prompts, streamlining data analytics processes.

Question Answering (QA) is one of the most prominent use cases of RAG, where users seek specific information in response to their questions. RAG systems retrieve relevant documents or knowledge snippets and use these to generate precise answers. This is especially useful when dealing with domain-specific or frequently updated information, such as medical databases, technical documentation, or customer support knowledge bases (Karpukhin et al., 2020). By combining the retrieval and generation components, RAG ensures that the generated answers are grounded in real-world information, making them both accurate and coherent.

Government entities have a variety of services offered to its users. These services range from granting licenses, generating reports, and issuing permits. The list of services goes on. As these services expand, end users get confused and they require a means to inquire about these services. In the past years, websites have served as the primary source of information to inform users about governmental service information. However, with the advent of chatbots technologies and LLM's, these technologies are increasingly used to provide a conversational solution that customers can interact with and provide useful and accessible information to users. Specifically, RAG is the vanguard of serving information from knowledge bases to customers in a conversational style.

The goal of this case study is to evaluate and choose the best advanced RAG approach improving a virtual assistant service designed for governmental services. This use case focuses on services within the medical domain, specifically involving permissions, licenses, and issuing reports. It also includes related services such as renewing and canceling documents, and handling inquiries about selling various medical products. Governmental services referenced in Appendix 1 will also be considered. The performance of advanced techniques is compared against a baseline derived from a basic RAG pipeline. The evaluation assesses the quality of both the retriever and the generator, with recommendations for improving the naive RAG pipeline to be presented to SESTEK[1]. SESTEK is an AI company that provides conversational AI Products. As a part of their offering to their enterprise customers, SESTEK provides Virtual Assistant Solution utilizing RAG to answer user queries from a knowledge base provided by the enterprise customers.

The current RAG pipeline utilized by SESTEK includes the following. A naive RAG pipeline utilizing Text-embedding-3-large from open AI alongside Chatgpt 4o for generation. The Chunks are configurable, but the default is 512 tokens with 128 overlap. The matching uses cosine similarity to return a configurable number of chunks. The current number of chunks is 5 for this use case. The current implementation suffers from hallucination in some queries. Another issue is that the pipeline might not generate the answer that is sought after as it either gets the wrong context to generate the answer from or it does not receive the correct answer. In this use case, we would like to measure its performance in questions that require reasoning or chain of thought mechanisms.

The evaluation relies on a set of metrics, including Context Relevance, Answer Relevance, and Groundedness, to assess the performance of the proposed RAG pipelines. To begin, an embedding model is chosen for the testing pipeline. Following this, the curated dataset from the MOHAP website was used to test both the embedding model and the complete RAG pipelines. This study aims at evaluating alternatives to the native RAG pipeline to be used in question answering scenarios in conversational AI products. The current workflow used for this use case utilizes native RAG with OpenAI API embedding models. Mainly " Text-embedding-3-large" is used for the indexing and retrieval of documents.

## RAG in QA:

### Advantages of Using RAG in QA

Utilizing RAG for QA offers several benefits. The retriever component allows access to a broader range of information than a standalone generative model, which might be limited to the data it was trained on. This makes RAG suitable for answering questions requiring niche or up-to-date knowledge. Additionally, RAG's ability to use retrieved documents as input to the generator ensures that the answers are not only accurate but also contextually rich, reducing the risk of hallucination, a common issue with generative models (Izacard & Grave, 2021).

[1] https://www.sestek.com/tr

## Challenges in RAG for QA:

Despite its strengths, RAG in QA also faces challenges. One of the primary issues is selecting the most relevant passages from retrieved documents, especially when the query's intent is nuanced. Moreover, while the retriever can find relevant documents, the generator may still produce answers that deviate from the retrieved context, particularly if the retrieval step fails to capture all relevant information. The balance between retrieval accuracy and generative capabilities also affects response latency, which is critical for real-time QA applications (Thakur et al., 2021).

One effective strategy is refining the retrieval step through more context-sensitive embeddings and feedback mechanisms. This involves training embeddings tailored to specific domains, which enhances the retriever's ability to understand the subtleties of domain-specific queries. By leveraging more contextually aware embeddings, the quality of the retrieved documents is significantly improved, ensuring that the generation phase has access to richer and more relevant data. This is particularly valuable for providing precise answers in specialized fields such as legal or medical QA (Karpukhin et al., 2020).

Another method combines traditional keyword-based search with modern neural retrieval approaches. By using a dual system that integrates dense vector-based retrieval with keyword-based models like BM25, the strengths of both are brought to bear. This approach ensures that the retriever can handle both nuanced semantic queries and those that require exact keyword matches. It is especially useful when dealing with complex questions that require understanding both the context of the words and their precise meanings (Thakur et al., 2021).

Adding additional context to document chunks during the embedding process can also enhance QA systems. This method involves enriching retrieved documents with supplementary information, such as metadata or summaries, before embedding them. By doing so, the retrieval mechanism is better able to capture the broader context of a document, which results in deeper insights during the generation phase. This enriched input allows the generation phase to produce more nuanced and accurate responses (Izacard & Grave, 2021).

Additionally, augmenting documents with synthetic queries and titles during the indexing process can be particularly effective. By creating synthetic queries and integrating them into the document's representation, the system is able to expand the contexts under which a document might be considered relevant. Indexing these augmented embeddings ensures that a broader range of user queries can find relevant matches, leading to more accurate retrievals. This approach is valuable in cases where the original document might not directly match user queries but still contains essential information (Lewis et al., 2020).

Another challenge lies in the inherent capabilities of the retriever and the generator models. Many encoders lack the robustness to catch semantic nuances in data properly and cause a poor encoding and poor retrieval as a result. Similarly, an LLM that hallucinates inherently can be difficult to predict and can lead to hallucination or sometimes contradicting responses.

By implementing these advanced strategies, QA systems can overcome many of the inherent challenges of retrieval and generation, leading to more effective and reliable answers across a variety of specialized domains.

Research Questions:

- Considering the following RAG approaches (Naive RAG, HyDE,Query Expansion, Corrective RAG, Hybrid RAG, and LLM-augmented retrieval), which one of them offers the best pipeline for our case study of Question Answering with RAG?
- Considering the following Embedding models (OpenAI text-embedding-3-large, Cohere Embed Large Multilingual 3, VoyageAI voyage-multilingual-2, Gemini Text-embedding-004) which commercial embedding model offers the most accurate retrieval?
- Considering the following LLM models (OpenAI 4o, OpenAI 4o mini, Claude 3.5 Sonnet, Claude 3.5 Haiku, Gemini 1.5 Flash, Gemini 1.5 Pro, Cohere Command R, and Cohere Command R plus), which  generator models offer the highest quality responses measured by the RAG Triad for the best performing Pipeline?

Dataset:

The dataset was retrieved from MOHAP Website by providing the list of links in appendix-1 with a web content crawler. The crawler returns the text content of the website. This content was cleaned from header and footer text existing in all the pages in the page list.

Dataset Card:

**Language**: English
**Type**: text (.txt)
**Source**: Website of Ministry of Health and Prevention (MOHAP) Services
**Scrape** Type: Text-content
**Summary**: The websites provided contain information about the services and the licenses offered by the Ministry of health in the UAE. The information consists of fees, required documents, duration, and requirements for services.

Descriptive Statistics:

Basic Metrics:

Table 1 - Basic Corpus Statistics

| Metric | Measure |
|---|---|
| Total characters: | 477,416 |
| Total characters (no spaces): | 408,065 |
| Total words: | 76,691 |
| Total sentences: | 1,929 |

Word Statistics:

Table 2 - Word level statistics for MOHAP data

| Metric | Measure |
|---|---|
| Vocabulary size | 2,972 |
| Lexical density | 3.88% |
| Type-token ratio (TTR) | 3.88% |
| Hapax legomena (Said Only once) | 1,194 (1.56%) |
| Average word length | 5.31 ± 3.47 |
| Stop words percentage | 28.76% |
| Total unique words | 3538 |
| Average chunk length (words) | 63.57 |
| Average chunk coherence | 0.28 |
| Average words per URL | 402.97 |

- **Vocabulary size (2,972)**: This indicates a moderately diverse vocabulary in the dataset. A larger vocabulary can be beneficial for capturing a wider range of concepts, but it also increases the complexity of the embedding space and may require more data for accurate retrieval.
- **Lexical density (3.88%) and Type-token ratio (TTR) (3.88%)**: These low percentages suggest that the text may contain a lot of repetition or common words, which can affect the distinctiveness of embeddings and potentially impact retrieval performance. A number of words heavily influence the lexical density which can be seen in the  figure 01

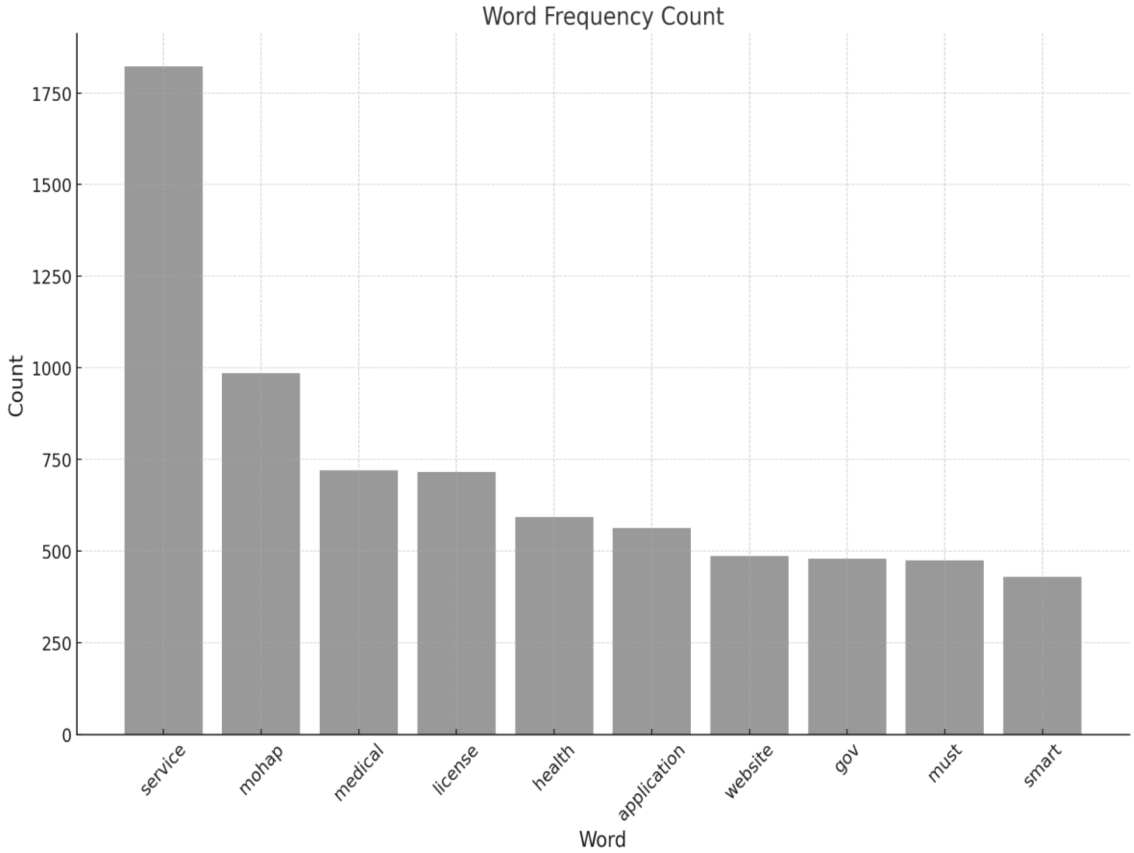

Figure 1 - Top 10 words in MOHAP corpus

Note that the sum of the percentages of the word frequency of each top-10 word to the total word count is about **10%.**

- **Hapax legomena (1,194, 1.56%)**: This shows that a significant number of words appear only once in the dataset. While some of these may be domain-specific terms, a high percentage could also indicate noise or inconsistency in the data. It might indicate typos or a significant number of unique words and concepts.
- **Average word length (5.31 ± 3.47)**: This suggests that the text uses a mix of shorter and longer words, which is typical for most documents.
- **Stop words percentage (28.76%)**: This is a standard percentage of stop words (common words like "the," "and," "a") in text.
- **Total unique words (3,538)**: This indicates the total number of distinct words in the dataset.
- **Average chunk length (words) (63.57)**: This is the average length of the text chunks used for embedding. Chunk size is crucial for balancing context and retrieval efficiency.

### Evaluation Questions and Ground Truths:

The Questions are a set of 5 questions curated by the ministry of health in addition to questions compiled by SESTEK team to test the current implementation and compare it to proposed pipelines. The questions were compiled by going to the website, creating the question and retrieving the answer from the related website. You can refer to Question Answer pairs in the table in appendix-2.

A set of multihop questions was created. These questions necessitate chain-of-thought reasoning to be answered. These questions followed a paper by Trivedi et al., 2022 that explained a top-down approach to creating multihop questions. By first creating a set of single-hop questions and linking questions that address the same entity we can construct such a question. For example, Table 3, shows how to use two questions to construct a multihop question. From the pool of single-hop questions we chose the following questions. “What is the section that is fee exempt from the fees of the application of a good standing certificate?” and “What are the requirements to apply for a good standing certificate for government Sector”. We notice that the answer of the first question is “Government Sector” which is a key entity in the second question. Note that there is no other hint to guide the LLM to answer the question from a single reasoning step. Then, these two questions are merged such that the key entity is replaced with information from the first question. The result is:

*"What are the requirements for the good standing certificate of medical staff in the sector that is fee-exempt for renewal staff licenses?"*

This question answers two parts, which are what is the sector that is fee exempt and then ask about required documents to apply for a good standing certificate for that sector.

Table 3 - Single Hop Questions used to create a Multihop question

| Single Hop Questions | Answer |
|---|---|
| What is the section that is fee exempt from the fees of the application of a good standing certificate? | **Government Sector** |
| What are the requirements to apply for a good standing certificate for **government Sector** | Letter of experience from the Department of Human Resources - Ministry of Health and Prevention.<br>Letter of experience from the employer attested by the medical director (technical director) - for hospital<br>employees.<br>●<br>Letter of experience from the employer (facility) accredited by the medical director of primary health care centers,<br>as per the location of the facility - for workers in primary health care centers.<br>●<br>● Certified copy of highest educational qualification.<br>● Copy of assessment certificate (issued by the Ministry of Health and Prevention).<br>● Copy of valid passport |

## Methodology:

To evaluate the pipeline and propose improvements to the output accuracy, we will start by evaluating the currently used embedding model and compare the retrieval accuracy of this model with other embeddings. The embeddings are stored in Pinecone, a vector database that is easy to use. It allows us to monitor and isolate the embeddings of each model and evaluate the results. After the evaluation, an embedding model will be chosen for this use case and it will be used in the rest of the tests.

### Phase 1:

The process starts with chunking and embedding the chunks in pinecone using the following models. OpenAI text-embedding-3-large (OpenAI, 2024), Cohere embed-multilingual-v3.0 (Cohere, 2023), VoyageAI Embedding MultiLingual Large 2 (Voyage AI, 2024), and Gemini text Embeddings 004 (Google AI, 2024). Each model has its own index in Pinecone's vector database. Then a set of 23 questions is used to retrieve related documents.

The outlined process, which involves chunking and embedding text using multiple language models (OpenAI, Cohere, VoyageAI, and Gemini) and storing them in a Pinecone vector database, is being used to create a robust system for document retrieval. This system is evaluated using metrics like NDCG@K,Recall@K, and IR Hit Rate. These metrics measure the effectiveness of the information retrieval system at returning relevant documents, with **K** representing the number of top results considered. NDCG@K focuses on the ranking of relevant documents, Recall@K assesses the completeness of the results, and IR Hit Rate indicates whether at least one relevant document is retrieved within the top **K** results. By using these metrics, the performance of the different embedding models and the overall document retrieval system can be comprehensively evaluated and compared.

### Phase 2:

The second phase starts by setting a baseline from the native RAG approach, also different Large language models are used in this phase to detect if the use of a large LLM is unnecessary for this use case, and to evaluate different commercial alternatives like Anthropic, Gemini or Cohere. This is a list of the models that we study:

- OpenAI 4o
- OpenAI 4o mini
- Claude 3.5 Sonnet
- Claude 3.5 Haiku
- Gemini 1.5 Flash
- Gemini 1.5 Pro
- Cohere Command R plus

It is worth mentioning that in some Advanced RAG systems, there is a need to use the LLM on multiple phases and the chosen model is used in the whole pipeline. The evaluation is based on the RAG Triad in addition to a mix of BLEU and ROUGE metrics for Queries and Outputted responses by the Large language model. By the end of this phase, an LLM is picked to use during our upcoming testing of the advanced rag pipelines comparison baseline is established with the native RAG implementation with the implementation utilizing OpenAI's Text embedding and GPT 4o.

### Phase 3:

The third phase consists of using the chosen LLM and the chosen embedding model in advanced RAG pipelines to compare with the RAG Triad metrics of the Native Rag approach with OpenAI models.

## Advanced RAG:

### Astute RAG:

Astute RAG is an approach that utilizes the following components. Internal knowledge which comes from the LLM used, and external knowledge which comes from the specified knowledge base (Website, or external corpora). An LLM is used to first filter and consolidate data from the external and internal databases. First, a query is transformed into an embedding and k-chunks are retrieved. Then, an LLM is used to generate internal knowledge related to the query  to account for missing information in the retrieved chunks or missing information in the internal knowledge base. An LLM is used to evaluate the knowledge retrieved and  consolidate information relevant to the query. The process of knowledge filtering and consolidation can be done iteratively if needed. The final step is relaying consolidated information to an LLM to generate a final response.

This approach enables the RAG pipeline to reason effectively over the knowledge retrieved from the external knowledge base, ensuring that the returned data is accurate, non-conflicting, and complete. As a result, the pipeline will operate with greater intelligence and reliability. (Wang et al., 2024)

### Self RAG:

Self-RAG enhances Retrieval-Augmented Generation (RAG) by incorporating dynamic retrieval, generation, and self-critique through "reflection tokens." The model decides if retrieval is needed, processes relevant documents, and generates outputs while evaluating relevance, factuality, and utility. This iterative process refines responses for accuracy and completeness. Unlike standard RAG, Self-RAG adapts retrieval and constraints dynamically, balancing creativity and factuality for diverse tasks (Asai et al., 2023).

For this project the evaluation LLM generated an improved question to retrieve ambiguous or missing information.

### Corrective RAG:

Corrective RAG is a robust approach designed to address inaccuracies in retrieval-augmented generation pipelines. It incorporates a lightweight retrieval evaluator to assess the quality of retrieved documents, assigning confidence levels—Correct, Incorrect, or Ambiguous—based on relevance. Correct results undergo refinement using a decompose-then-recompose strategy to extract critical information. Incorrect results trigger large-scale web searches to enhance data scope and accuracy. Ambiguous results combine both methods to ensure robustness.

This methodology enhances traditional RAG by improving knowledge accuracy and retrieval reliability. Experiments have demonstrated significant performance improvements across various datasets, showcasing its adaptability and effectiveness in mitigating retrieval errors while enhancing the robustness of generation tasks (Yan et al., 2024).

### Hybrid RAG:

This model has several differences from the native RAG model. Both dense retrieval and sparse retrieval are used to retrieve documents. Neural Embeddings are used as dense retrieval mechanisms where we get the K-nearest vectors to the query vector. BM25 is the sparse retrieval method. This method helps in getting the best 25 matching chunks in terms of normalized term frequencies. Normalization depends on chunk length. By combining both retrievals, we ensure that both the semantic meaning and the lexical representation of the queries are considered in relevant chunk retrieval.

BM25 is a ranking algorithm that utilizes the frequency of occurrence of words and the rarity of that word in the document (Chunk). The rarity is calculated by finding the inverse term frequency of the word in question. Then by combining these factors, a score is calculated which is then used to rank the chunks containing the words in the query.

### HyDE RAG:

HyDE (Hypothetical Document Embeddings) enhances Retrieval-Augmented Generation (RAG) by leveraging a Large Language Model (LLM) to generate hypothetical documents based on user queries. These hypothetical documents, written in the style of plausible responses, are used as embeddings to retrieve real, relevant documents from an external knowledge base.

The process begins with a query, for which the LLM generates a hypothetical response. These responses are called Hypothetical because we force the model to hallucinate answers without ground truth answers on retrieved data. A retriever then embeds and uses these generated documents to locate the most relevant real documents in the knowledge base. This combination ensures retrievals are grounded in actual data while filling knowledge gaps or ambiguities with contextually accurate hypothetical content (Gao et al., 2023).

By consolidating real documents with hypothetical ones, HyDE ensures responses are comprehensive, contextually rich, and grounded in non-conflicting knowledge.

### RAG with Query expansion:

Query Expansion with Generated Answers enhances Retrieval-Augmented Generation (RAG) by using a Large Language Model (LLM) to generate plausible answers based on the user's query. These generated answers act as enriched contextual documents, expanding the original query to improve retrieval performance. Unlike traditional query expansion methods, where related terms or rephrased queries are used, this approach directly creates hypothetical responses, simulating what an answer might look like. (deepset, n.d.).

The process begins with the user query, which is fed into an LLM to generate a hypothetical answer. These generated answers are treated as contextual documents and embedded into a vector space. The embeddings are then used to retrieve relevant real documents from the knowledge base, ensuring that the retrieval process is guided by contextually rich and detailed content. By embedding these hypothetical answers, the system fills in knowledge gaps and resolves ambiguities in the original query while still anchoring the final response in real, reliable data.

This approach ensures that RAG systems not only retrieve accurate information but also provide responses that are enriched with context and tailored to the user's intent, even when the initial query lacks specificity or completeness.

## RAG Evaluation:

### Evaluation Strategy:

This case study approaches assessment from two different sides. RAG Answer evaluation and retrieval evaluation.

For Embedding evaluation, NDCG@k is used with TruLens "trulens.feedback.groundtruth library", https://www.trulens.org/reference/trulens/feedback/groundtruth/ (*Groundtruth - □ TruLens*, n.d.). This class contains the following methods. Trulens is a tool that is used to facilitate and scale LLM evaluation. Traditionally, LLM evaluation is done by utilizing human experts to evaluate the output of Large language models. However, It is expensive to scale evaluation operations, to facilitate this process, Trulens employs LLM’s to evaluate LLM applications with its optimized prompts and evaluation pipelines. Trulens offers a wide range of feedback functions to evaluate different parts of the LLM application pipeline. Namely, pre-retrieval, post-retrieval, and generation of content by LLM.

### Ndcg_at_k:

To calculate NDCG@K, First, Calculate Discounted Cumulative Gain. Measures the relevance of items ranked in a list. Their position in the retrieved chunk rankings discounts this measure.

$$DCG@K = \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{log( i + 1 )}$$

Where K is the number of retrieved chunks and $rel$ is the relevance score for each retrieved chunk. The relevance score is measured based on the average of the lexical similarity by averaging ROUGE and BLEU score of each chunk.

Then the ideal DCG@K is calculated. It is the maximum possible DCG if the list of chunks is perfectly ranked.

$$IDCG@K \ = \sum_{i=1}^{K} \ \frac{2^{{rel_i}^{ideal}} - 1}{log(\ i + 1\ )}$$

Where ${rel_i}^{ideal}$ is the relevance score of a chunk i. If it is in the correct order. In short it calculates the relevance scores in descending order

NDCG@K is then calculated by dividing the calculated DCG@K to the ideal DCG@K (IDCG@K)

Formula 1

$$NDCG@K \ = \frac{DCG@K}{IDCG@K}$$

- Ground Truth answer (semantic similarity)

Trulens offers a variety of methods to measure ground truth similarity. BLEU, ROUGE and LLM based similarity calculation. However, Since the relevances is calculated by averaging BLEU and ROUGE for each chunk, and to capture the semantic similarity of the chunks; LLM based evaluation uses Trulens provides multiple methods for measuring ground truth similarity, including BLEU, ROUGE, and LLM-based similarity calculation. LLM-based evaluation is utilized to capture the semantic similarity of the chunks, since relevance is determined by averaging BLEU and ROUGE scores for each chunk.

- Recall_at_k

Formula 2

$$Recall@K \ = \frac{Number\ of\ relevant\ Items\ Retrieved\ in\ the\ K\ chunks\ retrieved\ from\ knowledgebase}{Total\ Number\ of\ relevant\ items\ in\ the\ dataset}$$

- Ir_hit_rate

It measures if at least one relevant item appears in the retrieved documents from the vector database. The output is a result of a binary function where:

Hit rate = 1 if at least one relevant document is within the K retrieved items

Or 0 otherwise.

Then Calculate the aggregate value as follows:

Formula 3

$$HitRate\ @K \ = \frac{1}{N}\sum_{q=1}^{N} \ \ HitRate$$

Where N is the number of queries.

These values are measured for the proposed Embedding models and it will be compared to OpenAI text-embedding-3-large.

When considering RAG output evaluation, the RAG Triad evaluation is conducted through Trulens to evaluate the output of the different RAG pipelines. The output of these experiments are kept and evaluated for each model. In this case study, we also compare the cost of inference of each model.

The indexing of the dataset is carried out using a recursive Text splitter from Langchain. The chunk size is determined as 512 and the overlap as 124. This goes for all data points in this experiment. Each experiment is done three times and the average of the RAG triad values are averaged to give us a single metric per RAG Pipeline. This step is repeated for all the LLM models we experimented with in each pipeline.

## Generation Evaluation

### Context Relevance:

This metric evaluates how well the submitted query aligns with the provided context or retrieved documents. This metric can be tested by either using Trulens evaluation metrics or by utilizing cosine similarity between the query and the retrieved chunks.

Relevance metrics are essential for evaluating how well information retrieval systems match user queries with relevant context and documents. Trulens Evaluation Metrics offer a comprehensive solution for assessing search result relevance, considering factors like semantic similarity and topical relevance to ensure that the retrieved information aligns with the user's needs.

### Answer Relevance:

This metric evaluates how directly the generated answer addresses the original input query. Even if the context is relevant and grounded, the answer must still provide useful information in response to the query.

The effectiveness of an answer lies not just in its factual accuracy and relevance to the given context, but also in its ability to directly address the user's query. Even if the provided information is grounded in reliable sources and pertains to the subject matter, it may still fall short if it does not specifically answer the question posed. This aspect of evaluation, which focuses on the directness and informativeness of the response in relation to the query, will be handled by Trulens. This ensures that the generated answers are not just contextually relevant and factually correct, but also effectively address the user's information needs.

### Groundedness:

It ensures that the generated answer remains faithful to the retrieved context. This is crucial as LLMs often expand beyond the facts present in the context, leading to hallucinated or exaggerated claims.

## Retrieval Evaluation

### Normalized Discounted Cumulative Gain at k (NDCG@k):

This metric evaluates the ranking quality of retrieved documents, emphasizing the importance of highly relevant documents being positioned higher in the list. Calculate the DCG@k for each query, then normalize it by the ideal DCG@k (IDCG@k). Average the NDCG@k values across all queries. Please refer to formula 1.

### Ground Truth answer (semantic similarity)

Ground truth, within the context of semantic similarity, refers to the human-curated benchmark or gold standard that represents the "true" or "ideal" similarity between items. These human judgments of similarity are typically based on expert knowledge or crowdsourced annotations.

### Recall at k (R@k):

This metric quantifies the ability to find all relevant documents within the top **k** retrieved documents. For each query, calculate the number of relevant documents in the top **k** and divide it by the total number of relevant documents for that query. Average the R@k values across all queries. Please refer to Formula 2

### Last Relevance Hit Rate (LR Hit Rate):

This metric checks if a relevant document is present within the last **k** retrieved documents. For each query, assign a 1 if a relevant document is found in the last **k** documents, and 0 otherwise. Average the values across all queries. Please Refer to Formula 3

# CHAPTER 2

## Experiments:

### Phase1: Retriever Evaluation:

As Discussed Before the following list of retrievers were tested against the 23 questions curated for this project  The goal is to evaluate the best retriever model for question answering utilizing MOHAP Dataset. You can refer to the questions in Appendix 2:
The following models are used in this experiment:

- Openai: Text-embedding-3-large
- Cohere Embed Large Multilingual 3
- VoyageAI voyage-multilingual-2
- Gemini Text-embedding-004

The test was done in two iterations. Iteration 1 focused on chunk size of 528 tokens and iteration 2 focused on 1000 tokens size. For each Iteration the testing included **k**  at 3 and 5. **k**  parameter here is the number of chunks retrieved per query from the vector database holding MOHAP chunks. This number was chosen as the  chunks related to ground truth was around 3-4 chunks per query.

To make a decision upon the best model, the following equation was agreed upon after talking to the product team of SESTEK. This equation was chosen to prioritize the semantic similarity to the ground truth at the current stage of the evaluation.

*Formula 04:*

$$0.1 * NDCG@k + 0 * 25 * IR\ Hit\ Ratio + 0.4 * Semantic\ Similarity + 0.25 * Recall@k$$

This gives more importance to Semantic similarity to ground truth and the Information retrieval hit ratio as they prioritize semantic similarity over ranking and recall of relevant chunks.

Trulens Feedback functions were used and ran over the ground truth questions for MOHAP dataset.

### Phase2: Evaluating LLM’s with Naive RAG:

The goal of this phase is to measure the baseline performance of Naive RAG with OpenAI generator and retriever. After that the retriever chosen from phase 1 is used with a list of LLM’s to test the performance of different generator models against the baseline. Testing is conducted for both direct questions within one category and multihop questions with yes-no responses within another category. The reason behind this is that it was noticed that LLM suffers particularly with these types of questions when compared to direct questions. Direct questions here cover single hop questions that start with what, where, when, and how. To account for the variability in LLM responses, the test is run three times for each response and the results are averaged. Note that this is done automatically within Trulens evaluation pipeline. Each question category is tested against the RAG triad (Context relevance, Answer Relevance, and Groundedness) by the end of this phase an LLM generator model is chosen to carry out phase 3.

### Phase3: RAG Approach Evaluation:

In this phase, the main expected output is evaluating a number of RAG pipelines using the retriever form phase 1 and the generator from phase 2 and deciding upon the best RAG approach and comparing them to the Naive RAG baseline established in phase 1.

## Discussion of the results:

### Phase 01: Retriever Evaluation:

The experiments with the aforementioned retriever models can be summarized in the following figures:

#### Ground Truth Similarity

Analysis of the retrieval system's performance reveals a clear trend based on chunk size. When utilizing a chunk size of 1000, the system achieves a significantly higher semantic similarity score for ground truth answers, averaging around 0.787 for the top performer (cohere @5). This is notably better than the 0.726 average achieved with a chunk size of 528 for the same model. Across all models and recall depths (k=3 and k=5), the larger chunk size consistently yields superior results, indicating that providing more context within each chunk enhances the retrieval system's ability to accurately identify semantically relevant information. In particular, we see a boost in performance across the board when increasing the chunk size, such as for the cohere @ 5 which increased from 0.726 to 0.786. For detailed results, refer to the figure 2 below

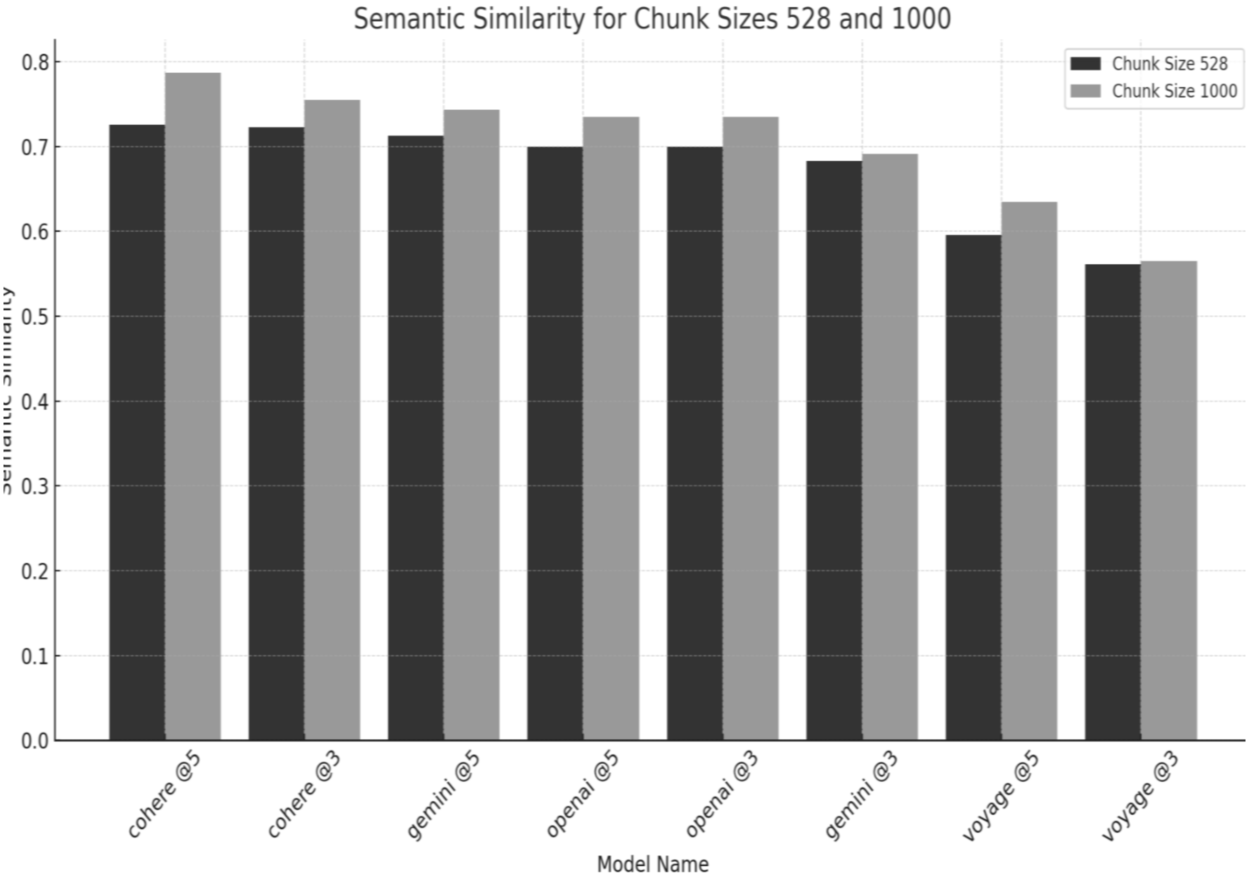

Figure 2 - A chart demonstrating semantic similarity to ground truth answers for chunk sizes 528 and 1000. The model was tested for k = 5 and k=3.

IR Hit Rate:

Switching to Information Retrieval (IR) hit rate, we again observe the impact of chunk size. With a chunk size of 1000, the system demonstrates a higher hit rate, exemplified by Openai @5 achieving 0.696. This contrasts with the performance at a chunk size of 528, where the same model (Openai @5) scores 0.652. Across the board, the larger chunk size consistently produces superior hit rates, suggesting that the increased contextual information within each chunk significantly improves the likelihood of retrieving at least one relevant document. Therefore, a chunk size of 1000 appears to be more effective for maximizing the retrieval of relevant information in this system. It's important to note that the voyage models tend to have a much lower hit rate, regardless of chunk size. Figure 3 demonstrates the results obtained from retrieval evaluation for information retrieval hit rate.

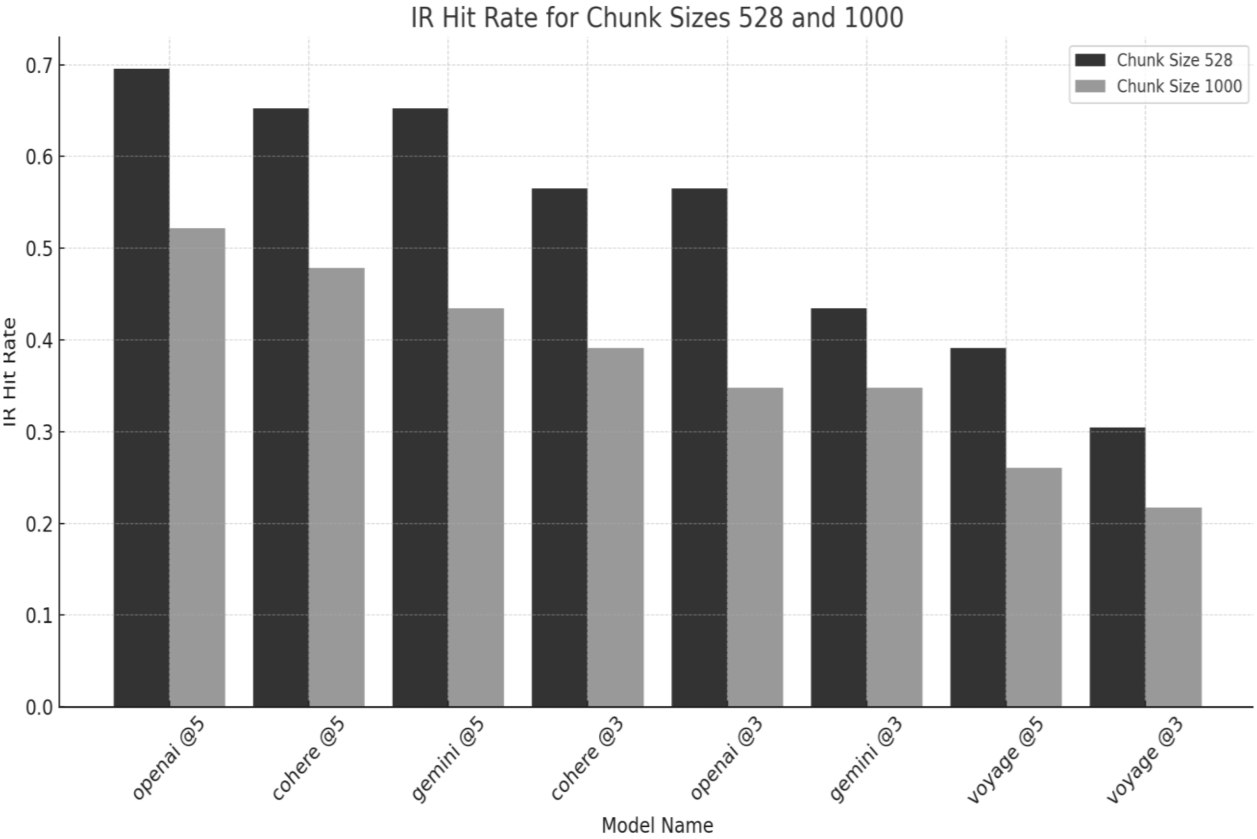

Figure 3 - A chart demonstrating information retrieval hit rate for chunk sizes 528 and 1000. The model was tested for k = 5 and k=3.

Recall @k

examining Recall@k at different chunk sizes (1000 and 500), we see a different pattern emerge compared to the previous metrics. At a chunk size of 500, the system generally achieves higher Recall@k scores. For instance, Openai @5 scores 0.255 with the smaller chunk size, compared to 0.131 with the larger chunk size of 1000. This suggests that while larger chunks excel at semantic similarity and hit rate, smaller chunks are more effective at retrieving a larger proportion of the total relevant documents. This likely stems from smaller chunks allowing for a more granular representation of the information space, increasing the chance of capturing at least a portion of a relevant document, even if it does not perfectly align with the query semantically. The voyage model consistently shows the lowest performance in terms of recall, further suggesting its relative weakness compared to the other models. This notion is supported by the nature of web pages containing these sets of information. Each url represents information about a single service. And each service page has small sections that represent different information about that service. For example fees, Required documents, audience etc. Hence, it is likely that smaller chunks have better recall @k.

The low score emphasizes the deficiency in context awareness within retrieval systems. This is due to the established methodology of ground truth systems, where, upon analyzing the question, all chunks containing pertinent contextual information are included. In numerous instances, the context is not explicitly stated within the question itself. Furthermore, this mirrors real-world scenarios. Figure 4 shows recall @k values for the retrieval evaluation on the Mohap data set.

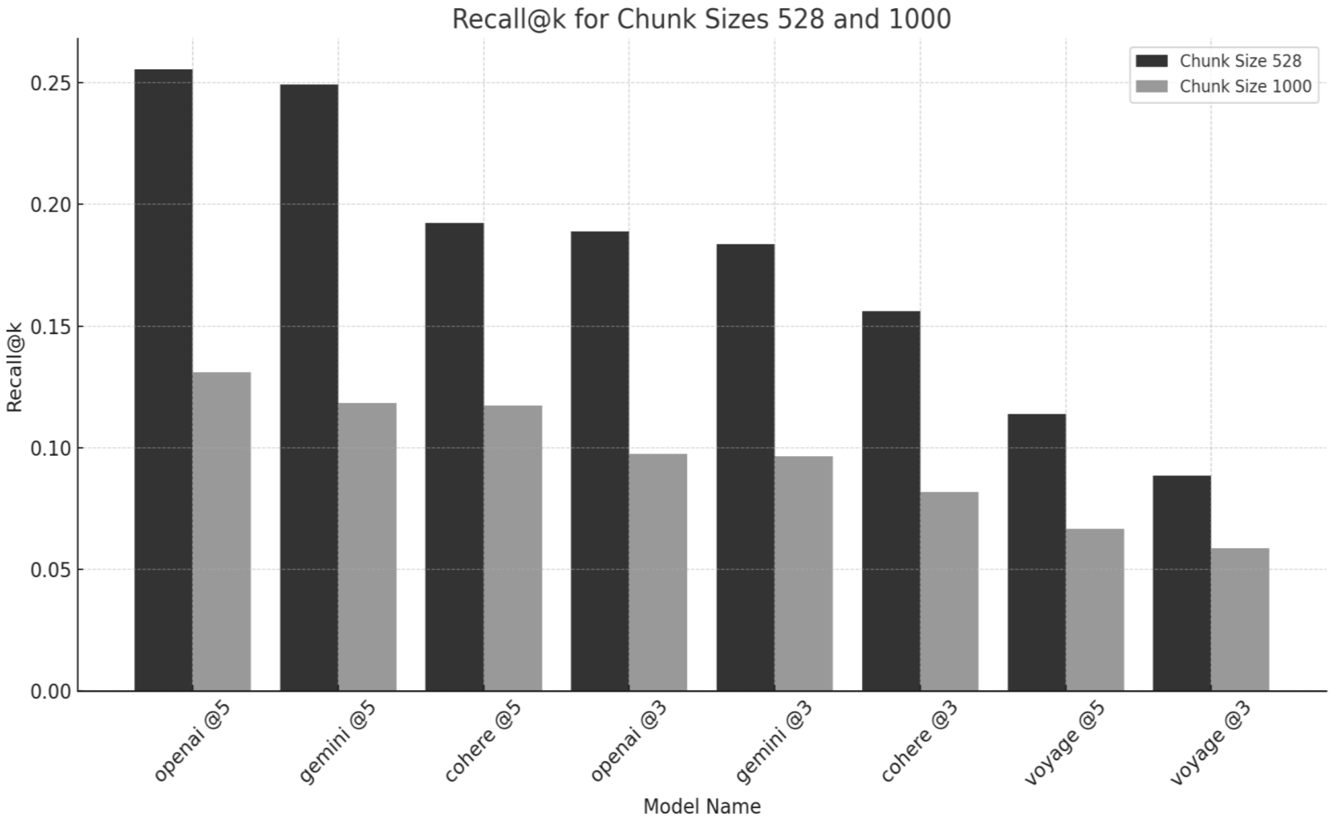

Figure 4 - A chart demonstrating recall at K-chunks for chunk sizes 528 and 1000. The model was tested for k = 5 and k=3.

### NDCG @K:

Finally, looking at Normalized Discounted Cumulative Gain (NDCG@k), we see perfect scores of 1.0 across all models and for both chunk sizes (500 and 1000). This indicates that the ranking quality of the retrieved documents is optimal in all tested configurations. In other words, the system is not only retrieving relevant documents but also consistently ranking the most relevant documents higher in the results list, regardless of chunk size or model used. This suggests that while chunk size impacts other aspects of retrieval performance, the ranking mechanism itself remains highly effective across different configurations.

### Conclusion:

chunk size significantly impacts retrieval performance. Smaller chunks (500) generally yield higher overall scores due to better Recall, while larger chunks (1000) excel in Semantic Similarity and Hit Rate.

Openai @5 and gemini @5 consistently lead across both sizes, with OpenAI @5 having a slight edge at 500. voyage models perform relatively weaker but show some improvement with larger chunks. The ideal chunk size depends on the specific application's priorities, but 500 appears optimal for a balanced performance based on the current metric weighting**.**

500's optimality, given the current metric weighting, stems from its superior Recall@k performance. While 1000 excels in identifying semantically similar and achieving higher hit rates, 500's smaller chunks lead to a more granular search, capturing a larger proportion of all relevant documents. This advantage in Recall outweighs the gains in Semantic Similarity and Hit Rate seen with 1000, resulting in higher overall scores when considering the weighted combination of all metrics (where recall is weighted 25%). Therefore, chunk size 500 is a better option. For the overall scores abiding by formula 04, you can refer to the following figure 5 and appendix -3:

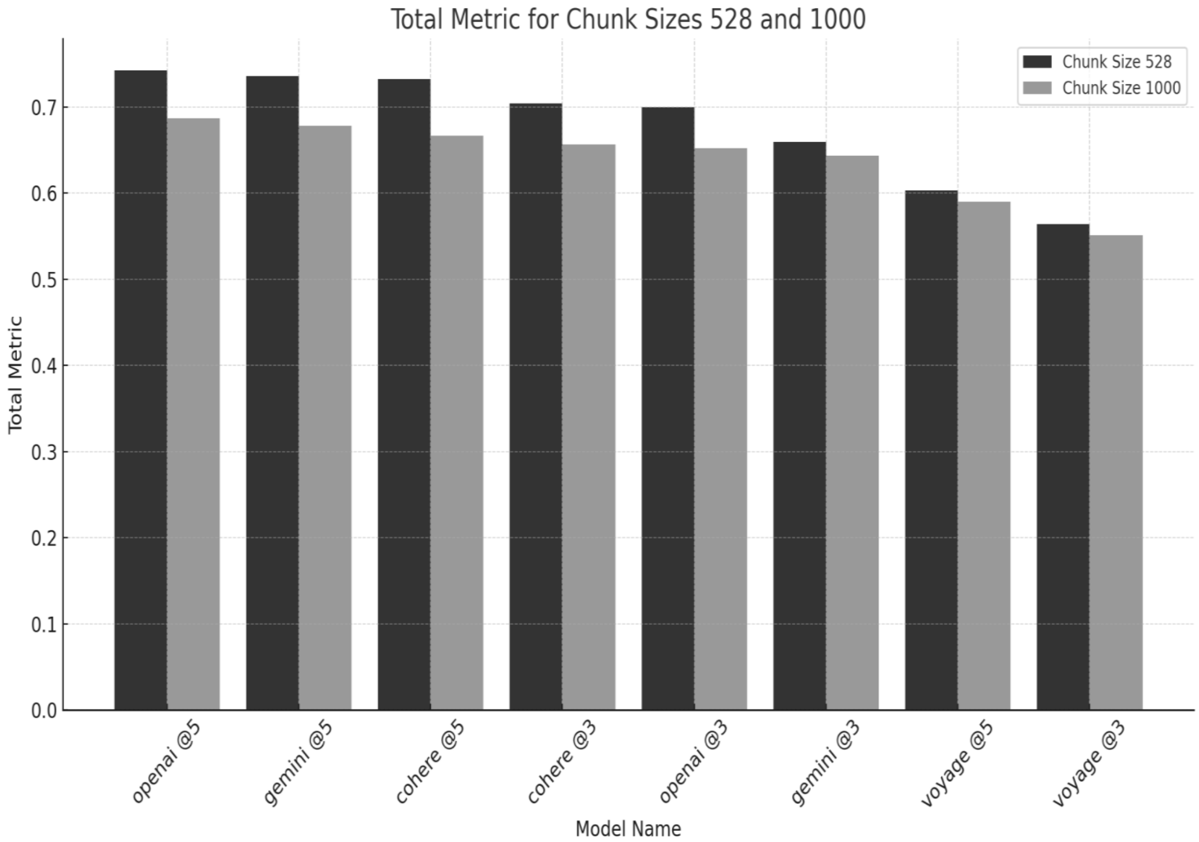

Figure 5 - A chart demonstrating the total score of the evaluation by following formula-4 for chunk sizes 528 and 1000. The model was tested for k = 5 and k=3.

Phase 2: Evaluating LLM's with Naive RAG:

Based on the results of phase 01, OpenAI text embedding large 03 was used in this experiment with 500-token sized chunk embeddings. Detailed experiment data can be found in appendix 4

The results are broken down in terms of the RAG Triad (Answer relevance, Context relevance, and Groundedness) and it is aggregated using the following:

*Formula 5*

$$0.4 * Answer\ Relevance\ +\ 0.2 * Context\ Relevance\ +\ 0.4 * Groundedness$$

Answer Relevance:

This evaluation highlights key differences in model performance across two question types. For complex, multi-hop yes/no questions, Claude 3.5 Sonnet, OpenAI's 4o, and Claude 3.5 Haiku, and OpenAI 4o mini lead with scores above 0.93, while Gemini 1.5 Pro lags. However, all models excel at direct questions, scoring 0.993 or higher, demonstrating near-perfect accuracy in basic information retrieval. This suggests a specialization among models, with some better suited for intricate reasoning while others excel at straightforward queries. Notably, Claude 3.5 Haiku, OpenAI 4o mini, and Command R Plus offer a strong balance of efficiency and accuracy across both question types.Figure 6 demonstrates the answer relevance scores for the generator models specified in the methodology.

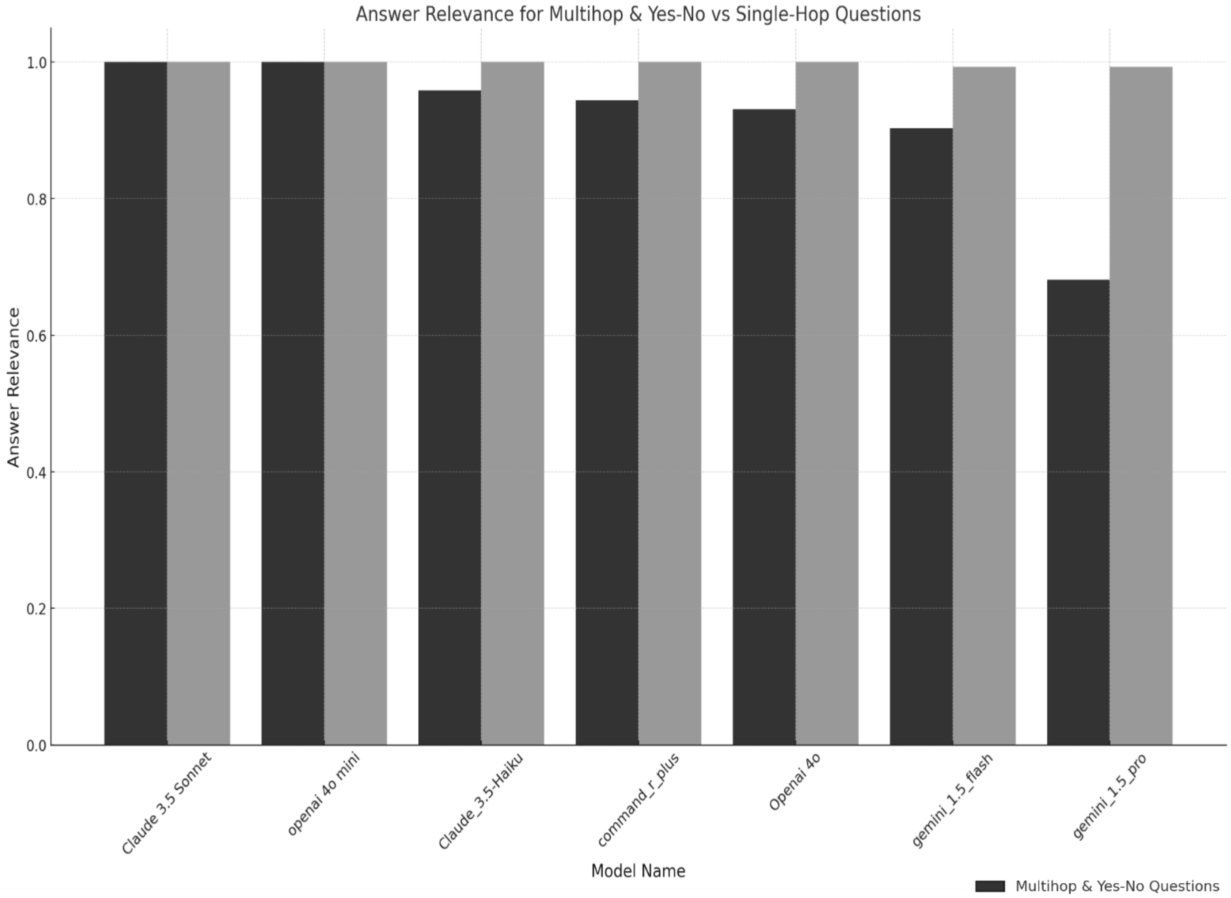

Figure 6 - A chart demonstrating Answer relevance for Multihop & Yes-No Questions and single hop questions.

### Context Relevance:

This data reveals the models' ability to provide relevant context when answering questions. For multi-hop yes/no questions, OpenAI 4o mini leads slightly with a score of 0.547, closely followed by Claude 3.5 Haiku and OpenAI 4o. All models perform relatively similarly in this category. In direct questions, Claude 3.5 Haiku takes the lead with a score of 0.745, showcasing a slightly stronger ability to offer relevant context. Overall, while there are minor differences, all models demonstrate a moderate ability to provide relevant context in their responses, with Claude 3.5 Haiku exhibiting a slight edge in direct questions and OpenAI 4o mini in complex ones. The scores in direct questions are higher than the scores in multi-hop.

To this end average chunk coherence was calculated for chunks of 528 tokens. The following equation was used:

Let $C = [C_1, C_2, \dots, Cn\ \ ]$ be the list of $n$ chunks. The coherence scores between consecutive chunks are represented as a list:

$$S = [S_1, S_2, \ldots, S_{n-1}\ \ ]$$

Where each coherence score $S_i$ is calculated as:

$$S_i \ = \ f(C_i, C_{i+1}\ \ )$$

for i=1,2,…,n−1 and $f(C_i, C_{i+1}\ \ )$ represents the semantic overlap function

$$f(C_i, C_{i+1}\ \ ) \ = \ \frac{|C_i \cap C_{i+1}|}{min\ (\ |C_i|\ , |C_{i+1}|\ )}$$

The resulting value was **0.284**: This score suggests that the chunks are poorly connected semantically. It implies that the chunks are not connected to each other and chunks do not flow into each other. This is expected to cause poor context relevance scores.

The low results especially for Multihop and Yes-No questions are also justified when considering the low chunk coherence mentioned in the data card and the low recall@k values for the retrievers. You can find a summary of the results in figure 7.

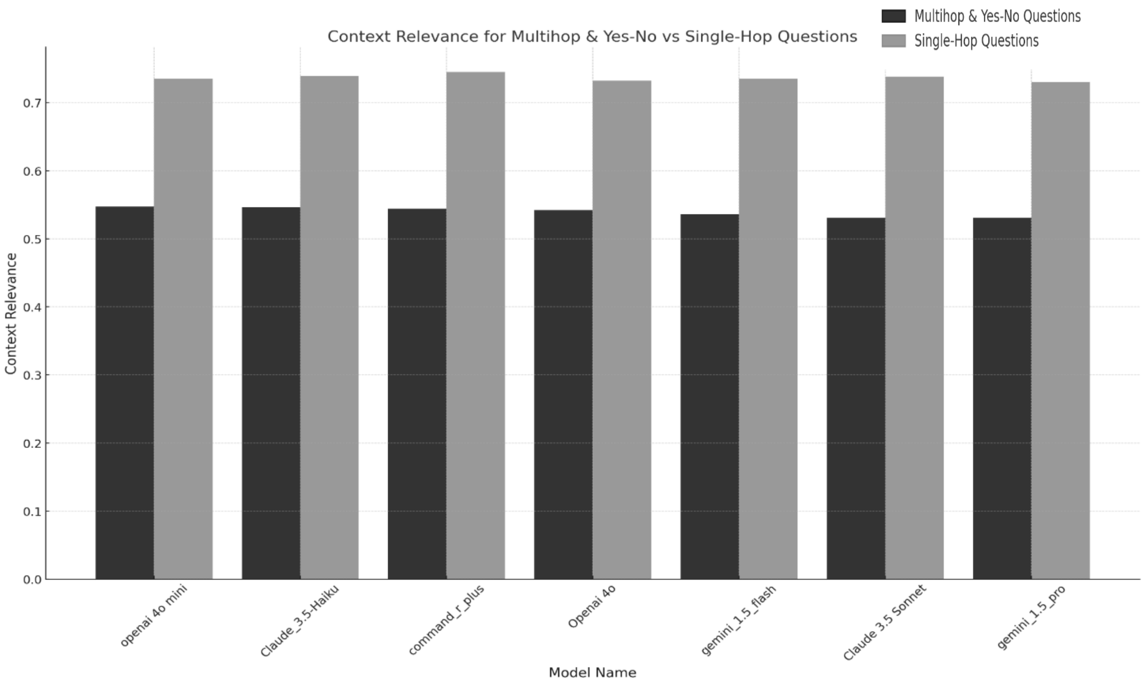

Figure 7 - A chart demonstrating context relevance for Multihop & Yes-No Questions and single hop questions.

### Groundedness:

The groundedness results in the context of RAG highlight strong performance for Gemini 1.5 Pro, which consistently achieves the highest groundedness across both multihop (0.968) and direct (0.975) scenarios. OpenAI 4o and Gemini 1.5 Flash also demonstrate robust groundedness, maintaining competitive scores in both categories. Notably, Claude 3.5 versions show slightly lower groundedness, with the Haiku variant outperforming Sonnet in multihop but trailing slightly in direct responses. OpenAI 4o Mini and Command R Plus consistently lag behind the top performers, indicating room for improvement in groundedness, particularly in multihop tasks. This underscores the value of advanced versions like Gemini Pro for tasks requiring high fidelity and contextual accuracy. Figure 8 shows groundedness scores for each generator model.

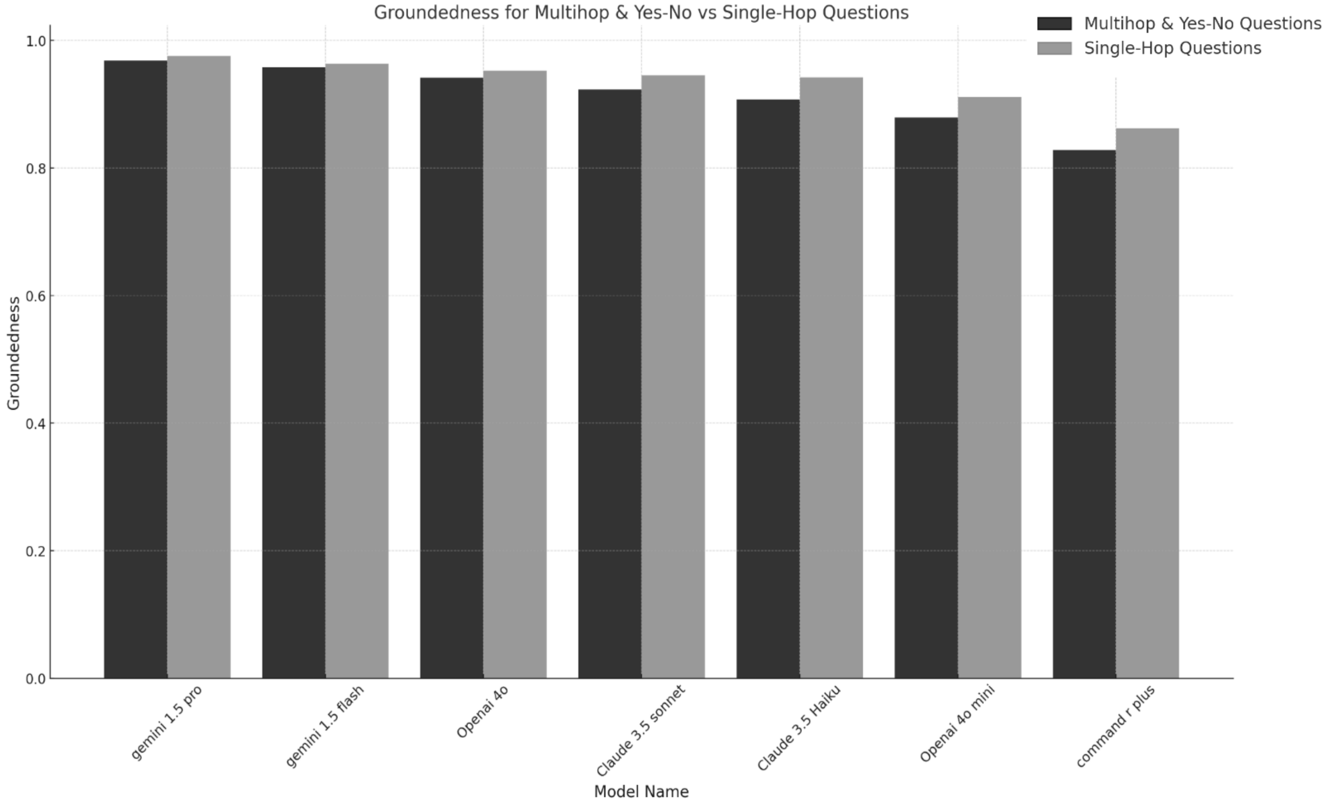

Figure 8 - A chart demonstrating groundedness scores for Multihop & Yes-No Questions and single hop questions.

Conclusion:

The weighted averages, combining answer relevance, context relevance, and groundedness, reveal interesting patterns for both multihop Yes-No and direct single-hop tasks.
For multihop Yes-No, Claude 3.5 Sonnet leads with the highest weighted average (0.869), closely followed by OpenAI 4o (0.864) and Claude 3.5 Haiku (0.8617). Gemini 1.5 Pro, while excelling in groundedness, scores relatively lower here (0.7658), indicating potential challenges in handling multihop Yes-No reasoning.

In direct single-hop, Gemini 1.5 Pro achieves the highest score (0.937), with Gemini 1.5 Flash (0.93) and OpenAI 4o (0.9286) following closely, demonstrating their strength in handling simpler, direct queries. Command R Plus, however, consistently scores lower across both tasks, suggesting it may lack the refinement needed for high relevance and groundedness.

Overall, Claude 3.5 Sonnet shines in multihop reasoning, while Gemini 1.5 Pro dominates direct single-hop tasks, emphasizing the importance of task-specific model selection. However, OpenAi 4o strikes a good balance between handling multihop and single hop questions It is going to be used for the rest of the experiments. Note that for Gemini 1.5 Pro, the experiment was done through Vertex AI API as opposed to Gemini 1.5 flash Which was done through Google AI Studio.

You can find a visual representation of the results in figure 9:

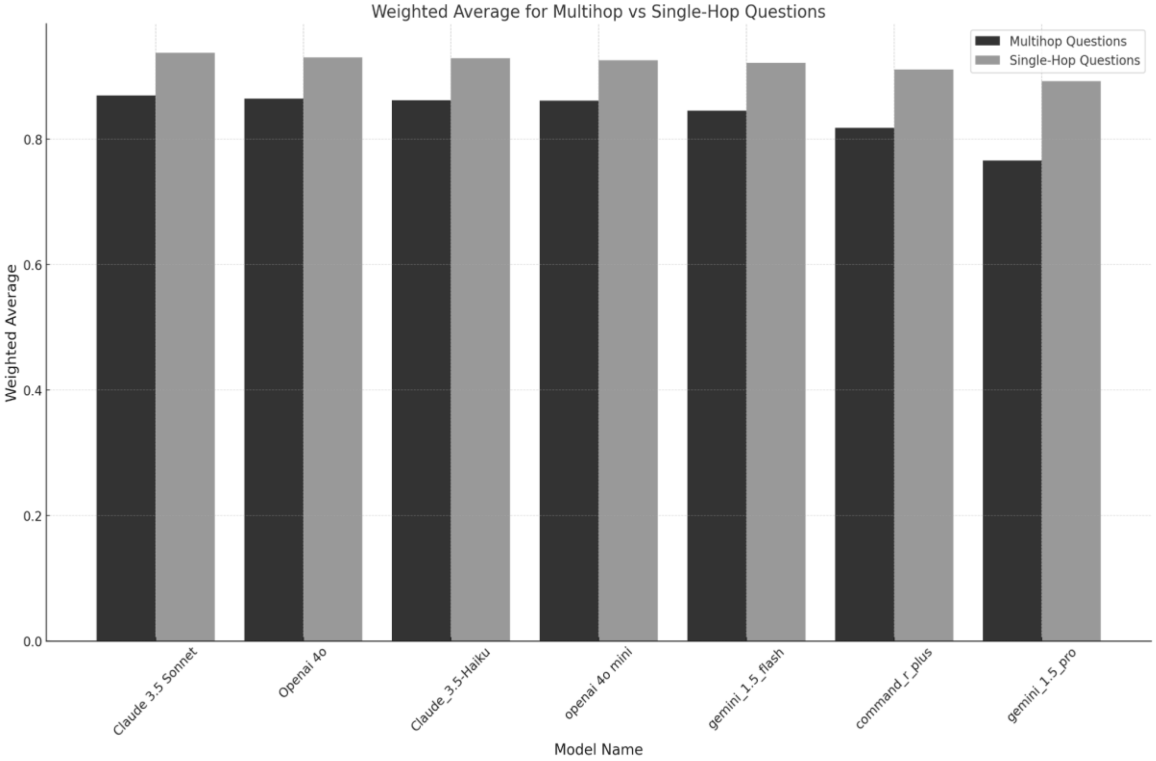

Figure 9 - A chart demonstrating aggregated value of the rag triad according to formula 5 for Multihop & Yes-No Questions and single hop questions.

## Phase 03: Evaluating LLM Approaches:

### Baseline Naive RAG Metrics:

By the end of the second phase, a baseline was established using OpenAI's retriever and 4o generator. The RAG triad results for this baseline are as shown in Table 4:

Table 4 - A table extracted from Trulens reporting page. It contains the baseline naive RAG scores.

| Generator Name | Retriever Name | Question Type | Answer Relevance | Groundedness | Context Relevance |
|---|---|---|---|---|---|
| Openai 4o | OpenAI Text Embedding Large 03 | Single hop | 1 | 0.952 | 0.739 |
| Openai 4o | OpenAI Text Embedding Large 03 | Multihop & Yes-No | 0.931 | 0.958 | 0.542 |

### Hybrid RAG:

To determine the optimal value for alpha in Hybrid RAG, initial testing was conducted. Alpha represents the respective contribution of the dense and sparse retrievers in the retrieval process. An alpha value of 1 signifies a completely semantic search, while an alpha of 0 indicates a keyword-based lexical search.

The initial testing revealed that the best performance was achieved with alpha = 1, suggesting that keyword contribution decreased accuracy. The balancing of the dense and sparse embedding values is managed through pinecone's hybrid_convex_scale function. It is hypothesized that the negative effect of the hybrid retrieval is attributed to either superior semantic search or the over-representation of the 10 most frequent words within the corpus.

For detailed results check appendix-5 and figure 10 below:

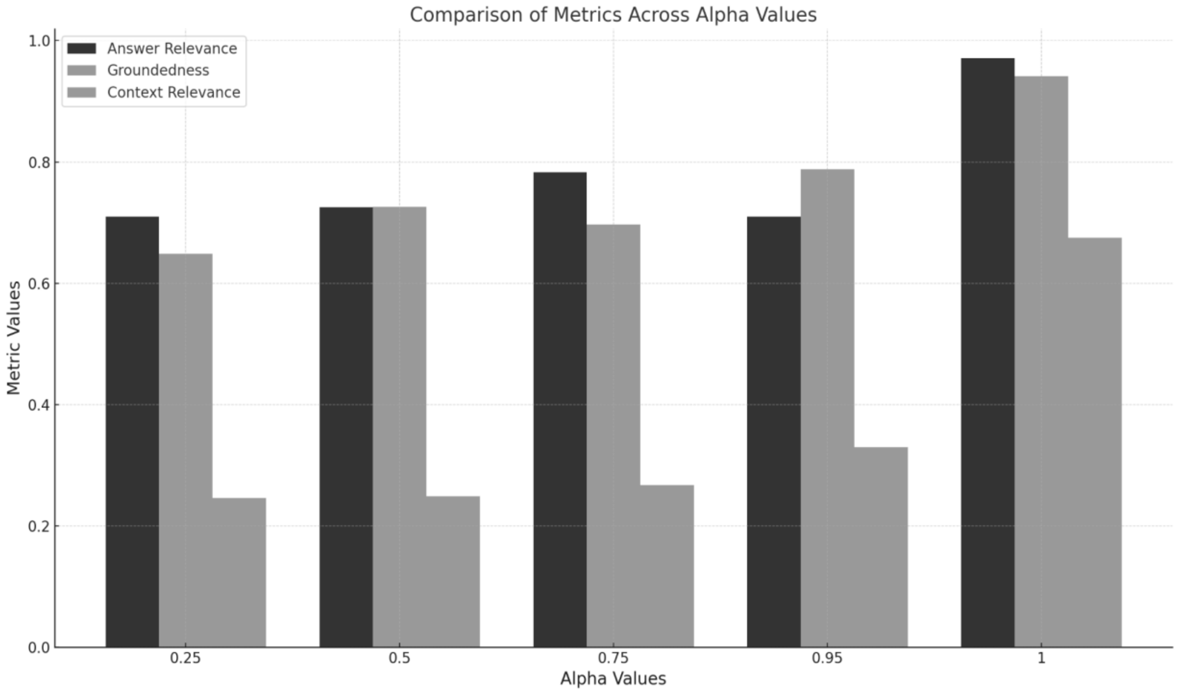

Figure 10 - A chart demonstrating Context relevance, Answer relevance, and groundedness scores for different alpha values in Hybrid RAG.

### Self-RAG:

Experimentation with Self RAG did not yield significant improvements. While it successfully identified single-hop questions as complete and refrained from triggering re-retrieval or regeneration, it failed to recognize the missing information in the chunks for most multi-hop questions. Only one question was re-generated; however, the LLM repeatedly rephrased the question with the word "Specifically." Despite attempting various prompting techniques to guide the model to generate a query that addressed the missing context, the most effective model was o1-mini when prompted with few-shot prompting that specifically addressed multi-hop reasoning. However, o1-mini also only detected an issue with one question. Another issue with o1 mini, is that it is not able to produce structured output to decide upon retrieval and regeneration of the answer based on the new query.

### Astute RAG:

Astute RAG utilizes internal LLM knowledge to generate context for the RAG system and can supplement the knowledge base with context. Testing was conducted on single-hop and multi-hop questions separately, with each test run three times. The system retrieved three chunks from the knowledge base and generated two chunks from the LLM for the k value.

As shown in table 5, Astute RAG's performance is significantly superior to naive RAG. Its weighted average score, based on Formula 5, is 0.8836 for multi-hop and yes-no questions and 0.966 for single-hop questions. In comparison, naive RAG scores 0.864 for multi-hop and yes-no questions and 0.929 for single-hop questions.

Astute RAG demonstrates notably better performance in context relevance, scoring 0.848 for single-hop questions and 0.606 for multi-hop questions. This is attributed to the more focused context generated from the LLM and the internal knowledge base. Thus, the three best-matching chunks and a list of two focused LLM-generated answers are submitted to an LLM to generate the final answer to the query.

Table 5 - A table presenting the RAG triad values for Astute RAG using Trulens.

| App Version | Question Type | Context Relevance | Groundedness | Answer Relevance |
|---|---|---|---|---|
| 4o-large_3-500-singlehop | single hop | 0.848 | 0.991 | 1 |
| 4o-large_3-500-Multihop Yes-No | Multihop & yes-no | 0.606 | 0.962 | 0.944 |

### Corrective RAG:

The results highlight the performance of Corrective RAG compared to Naive RAG across key metrics like groundedness, answer relevance, and context relevance for both multihop and single-hop question types is shown in table 6.

For multihop & yes-no questions, Corrective RAG achieves slightly higher groundedness (0.934 vs. 0.931) but slightly lower answer relevance (0.931 vs. 0.958) compared to Naive RAG. However, Corrective RAG significantly outperforms Naive RAG in context relevance (0.561 vs. 0.542), indicating that the corrective approach better aligns retrieved contexts with the generated answers in complex, multihop scenarios.

For single-hop questions, Corrective RAG shows slightly lower groundedness (0.958 vs. 0.952) but perfect answer relevance (1.0) compared to Naive RAG. Moreover, Corrective RAG outperforms Naive RAG in context relevance (0.831 vs. 0.739), demonstrating superior retrieval alignment for simpler queries.

Overall, Corrective RAG provides notable improvements in context relevance for both question types while maintaining competitive performance in groundedness and answer relevance. These results suggest that corrective measures enhance the retrieval-to-generation pipeline, particularly in aligning the context to the question's complexity.

Table 6 - A table presenting the RAG triad values for Corrective RAG using Trulens.

| App Version | Question Type | Context Relevance | Groundedness | Answer Relevance |
|---|---|---|---|---|
| 4o-large_3-500-singlehop | single hop | 0.831 | 0.958 | 1 |
| 4o-large_3-500-Multihop Yes-No | Multihop & yes-no | 0.561 | 0.934 | 0.931 |

RAG with Query Expansion:

The performance of RAG with Query Expansion and Naive RAG shows clear distinctions across various metrics. For multihop and yes-no questions, RAG with Query Expansion achieves a groundedness score of 0.903, lower than Naive RAG’s 0.958, indicating that Naive RAG provides more reliable and grounded responses. In terms of context relevance, RAG with Query Expansion scores 0.522, slightly below Naive RAG's 0.542, showing comparable but slightly weaker contextual understanding. For answer relevance, RAG with Query Expansion scores 0.833, which is also lower than Naive RAG’s 0.931, reflecting Naive RAG’s superior accuracy in handling complex questions.

For single-hop (simple) questions, RAG with Query Expansion performs better in context relevance, scoring 0.800 compared to Naive RAG’s 0.739, highlighting stronger contextual comprehension for simpler queries. Both systems exhibit similar performance in answer relevance, with RAG scoring 0.993 and Naive RAG scoring 1.0, showing nearly perfect accuracy for single-hop questions. In terms of groundedness, RAG with Query Expansion achieves 0.953, nearly identical to Naive RAG’s 0.952, indicating consistent reliability across both systems.

Overall, RAG with Query Expansion demonstrates better context relevance for single-hop questions but lags behind Naive RAG in groundedness and answer relevance, particularly for multihop queries. Naive RAG remains more grounded and accurate across both question types, making it more reliable for complex tasks.

Table 7 - A table presenting the RAG triad values for RAG with query expansion using Trulens.

| App Version | Question type | Groundedness | Context Relevance | Answer Relevance |
|---|---|---|---|---|
| 4o-large_3-500-Simple | singlehop | 0.953 | 0.8 | 0.993 |
| 4o-large_3-500-Multihop Yes-No | Multihop & yes-no | 0.903 | 0.522 | 0.833 |

HyDE:

HyDE and Naive RAG differ in their performance across various metrics and question types. For multihop and yes-no questions, HyDE achieves a context relevance score of 0.556, slightly better than Naive RAG's 0.542. However, Naive RAG outperforms HyDE in answer relevance, scoring 0.931 compared to HyDE's 0.903. In terms of groundedness, HyDE scores 0.918, which is slightly lower than Naive RAG's 0.958 for the same question type. This indicates that Naive RAG is better at providing more grounded answers for complex multihop questions, while HyDE struggles slightly in this area.

For single-hop questions, HyDE demonstrates stronger performance in context relevance, scoring 0.779, compared to Naive RAG's 0.739. Both systems achieve perfect answer relevance (1.0) for single-hop questions, indicating their ability to generate accurate answers. However, in terms of groundedness, HyDE slightly edges out Naive RAG with a score of 0.962 compared to Naive RAG's 0.952. This suggests that HyDE provides slightly more reliable answers in simpler scenarios where the question context is less complex.

The slight advantage of HyDE in some metrics is due to its ability to address the asymmetry caused by solely considering queries in semantic matching. Queries inherently contain aspects and entities of related chunks. However, this can lead to asymmetry in retrieval. To mitigate this, a hypothetical answer (essentially a hallucination) is generated to provide a more balanced semantic similarity for chunks in vector databases.

In summary, HyDE shows better performance in context relevance for single-hop questions but is slightly less grounded and relevant for multihop questions compared to Naive RAG. Naive RAG, on the other hand, consistently delivers strong performance in answer relevance and groundedness for both question types, making it a reliable option for complex queries.

Table 8 - A table presenting the RAG triad values for HyDE using Trulens.

| App Version | Question type | Context Relevance | Answer Relevance | Groundedness |
|---|---|---|---|---|
| 4o-large_3-500-Simple Questions | singlehop | 0.779 | 1 | 0.962 |
| 4o-large_3-500-Multihop Yes-No | Multihop & yes-no | 0.556 | 0.903 | 0.918 |

Conclusion:

In conclusion, the evaluation of various RAG systems demonstrates the nuanced trade-offs and strengths across different approaches. Naive RAG serves as a robust baseline, excelling in groundedness and answer relevance, particularly for complex multihop questions. Hybrid RAG, while introducing semantic and lexical balancing, underperformed due to over-representation of frequent words, highlighting the limitations of hybrid retrieval. Self RAG showed limited improvement, struggling to handle multi-hop questions effectively despite its attempts at query generation.

Astute RAG emerged as a strong contender, leveraging internal LLM knowledge to significantly enhance context relevance and overall performance, particularly for single-hop questions. Corrective RAG showcased improvements in context alignment while maintaining competitive metrics, indicating its effectiveness in refining the retrieval-to-generation process. RAG with Query Expansion demonstrated strengths in single-hop context relevance but fell short in groundedness and answer relevance for multihop questions. Lastly, HyDE exhibited better single-hop context relevance and groundedness but struggled with multihop questions due to inherent retrieval asymmetry.

Overall, the results emphasize the importance of tailoring RAG configurations to specific question types and highlight the potential for further innovations to optimize groundedness, relevance, and context alignment across complex scenarios.

A visual representation of the performance of different RAG approaches can be observed in the chart below, which adheres to formula 5.

*Multi-hop & Yes-No Questions*

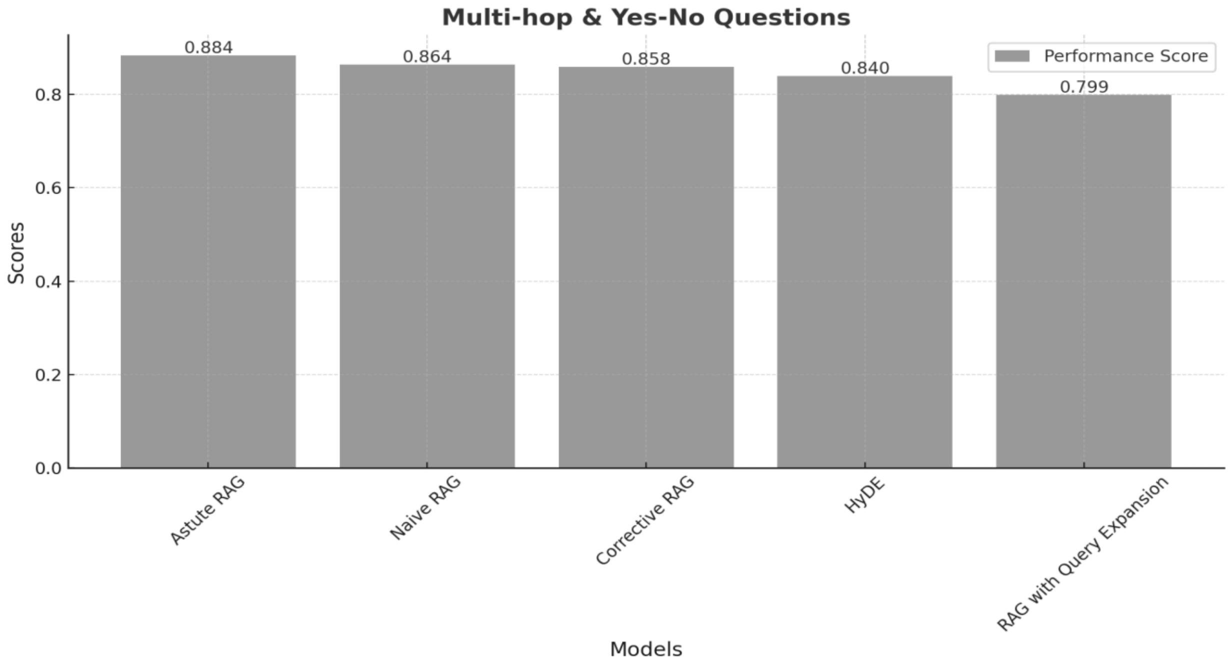

Figure 11 - A chart demonstrating The aggregated score for the RAG triad of multihop questions & Yes-no Questions according to formula 5.

*Simple Single-hop Questions*

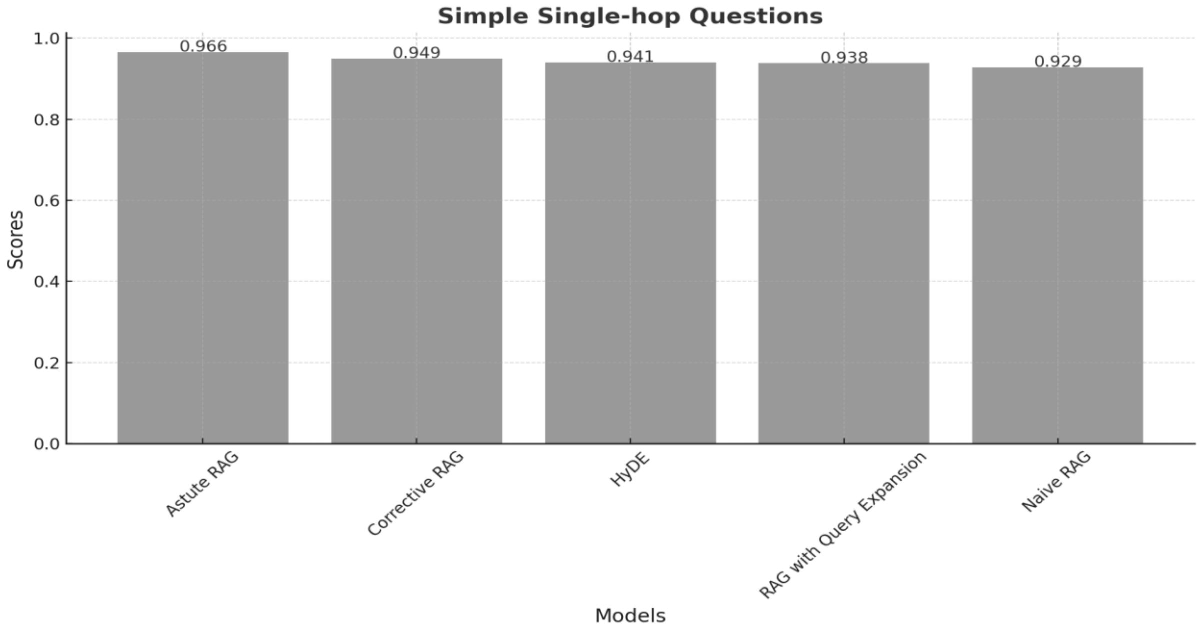

Figure 12 - A chart demonstrating the aggregated score for the RAG triad of Single Hop Questions according to formula 5.

# CHAPTER 3

## Discussion of the Results:

The experimental results across the three phases revealed several important insights about advanced RAG pipelines in the context of governmental services. In Phase 1, text-embedding-3-large from OpenAI emerged as the most effective retriever, particularly with a chunk size of 528 tokens, achieving superior balance across semantic similarity (0.700), IR hit rate (0.696), and recall@k (0.255). The choice of smaller chunks (528 tokens vs 1000) proved more effective for this specific use case, due to the modular nature of government service documentation where information is typically organized in discrete, topic-specific sections. As a part of future work discussions, the effect of plane crawling that ignores the hierarchy of the content in a URL is expected to have impacted the lower accuracy at lower chunk sizes.

Phase 2's evaluation of LLM generators demonstrated that while all models performed well on single-hop questions (scoring above 0.89 on the weighted average), there were significant variations in handling multi-hop queries. Claude 3.5 Sonnet led in multi-hop scenarios with a weighted average of 0.869, followed closely by OpenAI 4o (0.864), suggesting that more sophisticated models are particularly valuable for complex reasoning tasks. The notably lower context relevance scores across all models (averaging around 0.54 for multi-hop questions) highlight a persistent challenge in RAG systems: maintaining contextual coherence across multiple reasoning steps.

In Phase 3, Astute RAG emerged as the most promising advanced approach, achieving weighted averages of 0.966 for single-hop and 0.884 for multi-hop questions. This superior performance can be attributed to its effective integration of internal LLM knowledge with external sources, particularly beneficial for government service queries where both factual accuracy and comprehensive context are crucial. However, the increased computational overhead and complexity of implementation suggest a trade-off between performance and operational efficiency that organizations must consider.

Research Limitations:

This study is subject to several limitations that should be considered when interpreting the findings:

### Language Constraint:

The dataset and evaluations were conducted exclusively in English. This restricts the applicability of the findings to multilingual scenarios or contexts requiring language diversity, potentially limiting the generalizability of the results across different linguistic landscapes.

### Limited Chunk Sizes:

Only two chunk sizes (528 and 1000 tokens) were explored in this study. While these sizes were carefully selected, other chunking configurations might yield different insights, particularly for datasets with diverse structures and content.

### Evaluation Rounds:

The evaluation process assumes that three rounds of testing are sufficient to capture the performance variability of the models. This assumption may overlook subtleties in performance across additional iterations or under varying experimental conditions.

### Ground Truth Scoring Metrics:

Semantic similarity metrics such as BLEU and ROUGE were used to score the alignment between ground truth items and generated chunks. While these metrics provide valuable insights, they might not fully capture the nuanced relevance of certain responses, especially in complex or multihop question-answering scenarios.

### Simplified Evaluation Scenarios:

To facilitate scalability in web content evaluation, the study assumes that questions belong to a specific web page or set of related pages. The ground truth is defined as all chunks from these pages, which may not always align with real-world complexities, particularly for multihop questions requiring diverse or unrelated data sources.

### Data Collection Process:

Web scraping was conducted using an internally developed utility scraper, with data cleaning performed separately. This two-step approach might have introduced inconsistencies or overlooked nuances in the scraped data that could affect retrieval and generation quality.

### Chunk Overlap Issues:

The recursive text-splitting method caused significant overlaps in chunks for certain web pages, especially when smaller chunk sizes were used. This overlap might have introduced redundancy, impacting the effectiveness of the retriever and overall retrieval metrics.

### Default LLM Parameters:

Large language model (LLM) parameters, including temperature and penalties, were set to their defaults throughout the experiments. Specifically, a temperature of 1 was used for generated responses, which might not optimize performance for all scenarios or question types.

# CONCLUSION

This study demonstrates that while basic RAG systems can effectively handle straightforward queries about government services, more sophisticated approaches are needed for complex, multi-step reasoning tasks. The optimal configuration appears to be a combination of OpenAI's text-embedding-3-large retriever with 528-token chunks, coupled with either GPT-4 or Claude 3.5 Sonnet as the generator, implemented within an Astute RAG framework for complex queries.

These findings have important implications for deploying AI systems in government service contexts. While advanced RAG approaches like Astute RAG show superior performance, organizations must weigh the benefits against increased computational costs and implementation complexity. Future work should focus on optimizing context relevance scores, particularly for multi-hop queries, and exploring ways to reduce the computational overhead of advanced RAG approaches while maintaining their performance advantages.

For SESTEK and similar organizations implementing conversational AI in government services, these results suggest a tiered approach might be most effective: using simple RAG for straightforward queries while reserving advanced approaches for complex, multi-step questions. This would optimize both performance and resource utilization while maintaining high service quality across different types of user queries.

# FUTURE WORKS

Future works will focus on several dimensions. Parameter based experimentation with Naive RAG. This way, it will be possible to pinpoint the best configuration for better RAG performance. For example, doing experiments while controlling top-p and Temperature parameters and documenting the results for the changes in these parameters Another dimension is improving upon the crawling mechanism already implemented. The improvement should preserve the hierarchy of the webpage's html elements. This was proved true for the current implementation as the low scores of the smaller chunk sizes might be attributed to the fact that the crawling was simple.

As a part of future works, Arabic and Turkish language will be included in the evaluation. Additionally, more knowledge base sources will be included to encompass a wider range of use cases.

# APPENDICES

## APPENDIX-A

Table 9 - List of MOHAP Services and their corresponding web pages.

| Title | Link |
|---|---|
| Appeal Against Medical Licensing Committee Decisions | https://mohap.gov.ae/en/services/appeal-against-medical-licensing-committee-decisions |
| Request for a register for controlled or semi-controlled drugs custody | https://mohap.gov.ae/en/services/request-for-a-register-for-controlled-or-semi-controlled-drugs-custody |
| Classification of a product | https://mohap.gov.ae/en/services/classification-of-a-product |
| Renewal of Registration of a Conventional Pharmaceutical Product | https://mohap.gov.ae/en/services/renewal-of-registration-of-a-conventional-pharmaceutical-product |
| License for Nursing and Medical Professionals | https://mohap.gov.ae/en/services/license-for-nursing-and-medical-professionals |
| Renewal of a License to Practice as a Doctor | https://mohap.gov.ae/en/services/renewal-of-a-license-to-practice-as-a-doctor |
| Changing the Professional Titles of Nursing Practitioners and Medical Professionals | https://mohap.gov.ae/en/services/changing-the-professional-titles-of-nursing-practitioners-and-medical-professionals |
| Issue of Permit to Import Raw Materials | https://mohap.gov.ae/en/services/issue-of-permit-to-import-raw-materials |
| Changing the type of Medical activity of Private Health Facilities | https://mohap.gov.ae/en/services/changing-the-type-of-medical-activity-of-private-health-facilities |
| Licensing of a Pharmaceutical Facility | https://mohap.gov.ae/en/services/licensing-of-a-pharmaceutical-facility |
| Issue of a Certificate of Accreditation for a Center of Clinical Studies or Bioequivalence | https://mohap.gov.ae/en/services/issue-of-a-certificate-of-accreditation-for-a-center-of-clinical-studies-or-bioequivalence |
| Apply for Awareness or Educational Event | https://mohap.gov.ae/en/services/apply-for-awareness-or-educational-event |

| Title | Link |
|---|---|
| Request for a Price List of Registered Medications | https://mohap.gov.ae/en/services/request-for-a-price-list-of-registered-medications |
| Add Privilege for Health Professional | https://mohap.gov.ae/en/services/add-privilege-for-health-professional |
| Changing the name of private health facilities | https://mohap.gov.ae/en/services/changing-the-name-submit-the-service-application-online |
| Request to publish a research paper | https://mohap.gov.ae/en/services/طلب-نشر-ورقة-بحثية |
| Registration/ Renewal of Chemical Precursor Companies | https://mohap.gov.ae/en/services/renewal-of-chemical-precursor-company-registration |
| Re-license (Re-register) a Pharmaceutical Facility | https://mohap.gov.ae/en/services/re-licensing-a-pharmaceutical-institution |
| Renewal of Licenses for Nursing and Medical Professionals | https://mohap.gov.ae/en/services/renewal-of-licenses-for-nursing-and-medical-professionals |
| Renewal of Registration of a Pharmaceutical Product for General Sale | https://mohap.gov.ae/en/services/renewal-of-registration-of-a-pharmaceutical-product-for-general-sale |
| Renew a Health Facility License | https://mohap.gov.ae/en/services/renewal-of-a-private-medical-facility-license |
| Issue of Permit to Import Medical Equipment | https://mohap.gov.ae/en/services/issue-of-permit-to-import-medical-equipment |
| Issue/Cancel Pharmacy Home Delivery Permit | https://mohap.gov.ae/en/services/issue-cancel-online-pharmacy-permit-d996b0af |
| Issue an authorization to Import Narcotic Drugs | https://mohap.gov.ae/en/services/issue-of-permit-to-import-narcotic-drugs |
| Issue a Health Facility License | https://mohap.gov.ae/en/services/licensing-of-private-medical-facilities |
| Cancellation of a Pharmaceutical Facility's License | https://mohap.gov.ae/en/services/cancellation-of-a-pharmaceutical-facilitys-license |
| Re-license (Re-register) a Health Facility | https://mohap.gov.ae/en/services/re-issuance-of-a-private-medical-facilitys-license |
| Registration of a Change to the Professional Title of a Doctor | https://mohap.gov.ae/en/services/registration-of-a-change-to-the-professional-title-of-a-doctor |
| Changing the Licensed Ownership of Private Medical Facilities | https://mohap.gov.ae/en/services/changing-the-licensed-ownership-of-private-medical-facilities |
| Issue of a Certificate of Amendment of Registered Pharmaceutical Products | https://mohap.gov.ae/en/services/issue-of-a-certificate-of-amendment-of-registered-pharmaceutical-products |

| Title | Link |
|---|---|
| Complaints about private health facilities and their medical staff | https://mohap.gov.ae/en/services/complaints-about-private-health-facilities-and-their-medical-staff |
| Application to Change the Name of a Pharmaceutical Facility | https://mohap.gov.ae/en/services/application-to-change-the-name-of-a-pharmaceutical-facility |
| Re-licensing of Licenses for Nurses and Medical Professionals | https://mohap.gov.ae/en/services/re-licensing-of-licenses-for-nurses-and-medical-professionals |
| Request for a List of Licensed Pharmaceutical Facilities in the UAE | https://mohap.gov.ae/en/services/request-for-a-list-of-licensed-pharmaceutical-facilities-in-the-uae |
| Renewal of Registration Certificate for Medical Professional | https://mohap.gov.ae/en/services/renewal-of-registration-certificate-to-practice-nursing-and-midwifery |
| Transfer of a Doctor's License | https://mohap.gov.ae/en/services/transfer-of-a-doctors-license |
| Issue of Permit to Export Precursors Chemicals | https://mohap.gov.ae/en/services/issue-of-permit-to-export-precursors-chemicals |
| Transfer License for practicing the profession of pharmacist | https://mohap.gov.ae/en/services/transfer-license-for-practicing-the-profession-of-pharmacist |
| Issue of Permit to Import Medicines for Exhibitions | https://mohap.gov.ae/en/services/issue-of-permit-to-import-medicines-for-exhibitions |
| Appealing Against Health Advertisement Violation | https://mohap.gov.ae/en/services/appealing-against-health-advertisement-violation |
| Licensing of a Visiting Doctor from the UAE | https://mohap.gov.ae/en/services/licensing-of-a-visiting-doctor-from-the-uae |
| Issue of Permit to Import Chemical Precursors | https://mohap.gov.ae/en/services/issue-of-permit-to-import-chemical-precursors |
| Renewal of a Pharmaceutical Facility License | https://mohap.gov.ae/en/services/renewal-of-a-pharmaceutical-facility-license |
| Licensing of a Visiting Foreign Doctor | https://mohap.gov.ae/en/services/licensing-of-a-visiting-foreign-doctor |
| Registration of a Conventional Pharmaceutical Product | https://mohap.gov.ae/en/services/registration-of-a-conventional-pharmaceutical-product |
| Transfer of Licenses for Nurses and Medical Professionals | https://mohap.gov.ae/en/services/transfer-of-licenses-for-nurses-and-medical-professionals |
| Health Professional Evaluation | https://mohap.gov.ae/en/services/health-professional-evaluation |
| License for Healthcare Advertisement for a Non-Healthcare Institution | https://mohap.gov.ae/en/services/license-for-healthcare-advertisement-for-a-non-healthcare- |

| Title | Link |
|---|---|
|  | institution |
| Adding a Partner in a Private Medical Facility | https://mohap.gov.ae/en/services/adding-a-partner-in-a-private-medical-facility |
| Adding a New Specialty to Private Health Facilities | https://mohap.gov.ae/en/services/adding-a-new-specialty-to-private-health-facilities |
| Verification of Registration of a Nursing Certificate Issued Outside UAE | https://mohap.gov.ae/en/services/verification-of-registration-of-a-nursing-certificate-from-a-foreign-country |
| Changing the Technical Director of a Private Medical Facility | https://mohap.gov.ae/en/services/changing-the-technical-director-of-a-private-medical-facility |
| Application to Change the Location of a Pharmaceutical Facility | https://mohap.gov.ae/en/services/application-to-change-the-location-of-a-pharmaceutical-facility |
| Reporting adverse events of medical products from medical companies | https://mohap.gov.ae/en/services/evaluation-of-mandatory-reports-of-adverse-localized-negative-reactions-to-drugs |
| Provide controlled drugs prescription book | https://mohap.gov.ae/en/services/provide-controlled-drugs-prescription-book |
| Renew License of healthcare advertisement for a Licensed Healthcare Institution in the UAE | https://mohap.gov.ae/en/services/renew-license-of-healthcare-advertisement-for-a-licensed-healthcare-institution-in-the-uae |
| Apply for Healthy Restaurant Accreditation | https://mohap.gov.ae/en/services/apply-for-healthy-restaurant-accreditation |
| Issue a Good Standing Certificate for a Health Professional | https://mohap.gov.ae/en/services/issue-of-good-professional-conduct-certificates-for-health-professional |
| Approve narcotic drugs for internal pharmacies at private hospitals | https://mohap.gov.ae/en/services/approve-narcotic-drugs-for-internal-pharmacies-at-private-hospitals |
| Cancel a Health Facility License | https://mohap.gov.ae/en/services/cancelling-the-license-of-a-private-medical-facility |
| Issue/ Cancel Online Pharmacy Permit | https://mohap.gov.ae/en/services/issue-cancel-online-pharmacy-permit |
| Issue of permit to Export Medicines | https://mohap.gov.ae/en/services/issue-of-permit-to-export-medicines |
| Cancellation of a Doctor's License | https://mohap.gov.ae/en/services/cancellation-of-a-doctors-license |

| Title | Link |
|---|---|
| Cancellation of Licenses for Nurses and Medical Professionals | https://mohap.gov.ae/en/services/cancellation-of-licenses-for-nurses-and-medical-professionals |
| Renewal of a Healthcare Advertisement License for non Healthcare Institution | https://mohap.gov.ae/en/services/renewal-of-a-healthcare-advertisement-license-for-non-healthcare-institution |
| Analyze/ Re-Analysis of a Medical Product for A Pharmaceutical Company and Its Subsidiaries | https://mohap.gov.ae/en/services/analyze-medical-product-for-a-pharmaceutical-company-and-its-subsidiaries |
| Cancellation of a License to Practice as a Pharmacist | https://mohap.gov.ae/en/services/cancellation-of-a-license-to-practice-as-a-pharmacist |
| Changing the Location of a Private Medical Facility | https://mohap.gov.ae/en/services/changing-the-location-of-a-private-medical-facility |
| Issue of Permit to Import Medicines from a Local Agent | https://mohap.gov.ae/en/services/issue-of-permit-to-import-medicines-from-a-local-agent |
| Licensing of a Doctor | https://mohap.gov.ae/en/services/licensing-of-a-doctor |
| Renewal of a License for a Healthcare Advertisement in a Medical or Commercial Directory | https://mohap.gov.ae/en/services/renewal-of-a-license-for-a-healthcare-advertisement-in-a-medical-or-commercial-directory |
| Attestation of Medical Leaves and Reports | https://mohap.gov.ae/en/services/attestation-of-medical-leaves-and-reports |
| Issue of Permit to Import Medicines for Personal Use | https://mohap.gov.ae/en/services/issue-of-permit-to-import-medicines-for-personal-use |
| Approval of Pharmacovigilance Officer within UAE | https://mohap.gov.ae/en/services/approval-of-pharmacovigilance-officer-of-pharmaceutical-facility |
| Registration of a Medical Equipment | https://mohap.gov.ae/en/services/registration-of-a-medical-equipment |
| Registration of Pharmaceutical Product for General Sale | https://mohap.gov.ae/en/services/registration-of-pharmaceutical-product-for-general-sale |
| Renewal of the Registration of a Manufacturer of Medical Products | https://mohap.gov.ae/en/services/renewal-of-the-registration-of-a-manufacturer-of-medical-products |
| Issue of a Certificate of a Pharmaceutical Product for Export | https://mohap.gov.ae/en/services/issue-of-a-certificate-of-a-pharmaceutical-product-for-export |
| Registration of A Manufacturer of Medical Products | https://mohap.gov.ae/en/services/registration-of-a-manufacturer-of-medical-products |

| Title | Link |
|---|---|
| Renewal of Registration of Medical Equipment | https://mohap.gov.ae/en/services/renewal-of-registration-of-medical-equipment |
| Ajr Wa Aafya | https://mohap.gov.ae/en/services/ajr-wa-aafya |
| Renewal License for a Healthcare Advertisement Related to a Healthcare Event | https://mohap.gov.ae/en/services/renewal-of-license-of-a-healthcare-advertisement-related-to-healthcare-event |
| Licensing of a Doctor or Dentist to Utilize the Services of a Private Health Facility | https://mohap.gov.ae/en/services/licensing-of-a-doctor-or-dentist-to-utilize-the-services-of-a-private-health-facility |
| License for a Healthcare Advertisement on a Website or Digital Link | https://mohap.gov.ae/en/services/license-for-a-healthcare-advertisement-on-a-website-or-digital-link |
| Request a Register of the Narcotic drugs distributed to Private Health Establishments | https://mohap.gov.ae/en/services/request-a-register-of-the-narcotic-drugs-distributed-to-private-health-establishments |
| License for Healthcare Advertisement on Social Media | https://mohap.gov.ae/en/services/license-for-healthcare-advertisement-on-social-media |
| Request narcotics custody for hospital | https://mohap.gov.ae/en/services/request-narcotics-custody-for-hospital |
| License for a Healthcare Advertisement through Call Centers | https://mohap.gov.ae/en/services/license-for-a-healthcare-advertisement-through-call-centers |
| Approve Emergency Drugs and Psychotropic Materials | https://mohap.gov.ae/en/services/approve-emergency-drugs-and-psychotropic-materials-for-hospitals |
| Registration of A Pharmaceutical Product Derived from Natural Sources | https://mohap.gov.ae/en/services/registration-of-a-pharmaceutical-product-derived-from-natural-sources |
| Renewal License for a Healthcare Advertisement through Call Centers | https://mohap.gov.ae/en/services/renewal-of-license-for-healthcare-advertising-through-call-centers |
| Approve the Signature of the Responsible for Narcotic Drug Custody | https://mohap.gov.ae/en/services/approve-pharmacist-signature-for-narcotics-custody |
| Renewal of a Healthcare Advertisement on a Website or Digital Link | https://mohap.gov.ae/en/services/renewal-of-license-for-a-healthcare-advertisement-on-a-website-or-digital-link |
| License for a Healthcare Program | https://mohap.gov.ae/en/services/license-for-a-healthcare-program |

| Title | Link |
|---|---|
| Approve narcotic drugs for one day surgery centers | https://mohap.gov.ae/en/services/approve-narcotic-drugs-for-one-day-surgery-centers |
| Issue of a Certificate of Free Sale of a Medical Product for Export | https://mohap.gov.ae/en/services/issue-of-a-certificate-of-free-sale-of-a-medical-product-for-export |
| Reporting Side Effects of Medicines and Medical Products Within UAE for Individuals | https://mohap.gov.ae/en/services/reporting-side-effects-of-medicines-and-medical-products |
| Issue of a License for a Healthcare Advertisement in a Medical or Commercial Directory | https://mohap.gov.ae/en/services/issue-of-a-license-for-a-healthcare-advertisement-in-a-medical-or-commercial-directory |
| Issue a Price List for Medical Products Registered in the UAE | https://mohap.gov.ae/en/services/request-for-a-price-list-for-medical-products-registered-in-the-uae |
| Amend the Registration Data of a Medical Company or a Manufacturer Licensed to Market | https://mohap.gov.ae/en/services/issue-of-a-certificate-to-amend-the-registration-data-of-a-medical-company-or-a-manufacturer |
| Re-licensing of a Doctor | https://mohap.gov.ae/en/services/re-licensing-of-a-doctor |
| Issue a price list for controlled or semi-controlled medicines | https://mohap.gov.ae/en/services/request-for-price-list-for-controlled-or-semi-controlled-medicines |
| Issue of a Single Medical Product Pricing Certificate | https://mohap.gov.ae/en/services/issue-of-a-single-medical-product-pricing-certificate |
| Re-Pricing of a Medicine | https://mohap.gov.ae/en/services/re-pricing-of-a-single-medical-product |
| License for a Healthcare Advertisement Related to a Healthcare Event | https://mohap.gov.ae/en/services/license-for-a-healthcare-advertisement-related-to-a-healthcare-event |
| Issue a Price List of Medicines Registered to a Company | https://mohap.gov.ae/en/services/request-for-a-price-list-of-medicines-registered-to-a-company |
| Re-licensing of a Pharmacist | https://mohap.gov.ae/en/services/re-licensing-of-a-pharmacist |
| Renewal of a License to Practice as a Pharmacist | https://mohap.gov.ae/en/services/renewal-of-a-license-to-practice-as-a-pharmacist |
| Request Handover of Narcotic Drugs Custody | https://mohap.gov.ae/en/services/request-handover-of-narcotics-custody-among-pharmacists |
| Registration of a Change to the Professional Title of a Pharmacist | https://mohap.gov.ae/en/services/registration-of-a-change-to-the-professional-title-of-a-pharmacist |

| Title | Link |
|---|---|
| Authorize a Pharmaceutical Company to dispose any Pharmaceutical Products as Medical Waste | https://mohap.gov.ae/en/services/authorize-a-pharmaceutical-company-to-dispose-of-any-pharmaceutical-products-as-medical-waste |
| Renewal of Registration of a Pharmaceutical Product derived from Natural Sources | https://mohap.gov.ae/en/services/renewal-of-registration-of-a-pharmaceutical-product-derived-from-natural-sources |
| License of healthcare advertisement for a Licensed Healthcare Institution in the UAE | https://mohap.gov.ae/en/services/license-of-healthcare-advertisement-for-a-licensed-healthcare-institution-in-the-uae |
| Renewal License for Healthcare Advertisement on Social Media | https://mohap.gov.ae/en/services/renewal-of-license-for-healthcare-advertisement-on-social-media |
| Assessment of Medical Products for Pharmacological Research and Clinical Studies of Drugs | https://mohap.gov.ae/en/services/assessment-of-medical-products-for-pharmacological-research-and-clinical-studies-of-drugs |
| Renewal of a Healthcare Program | https://mohap.gov.ae/en/services/renewal-of-healthcare-program-license |
| Issue a Certificate of Compliance with the good Practice Standards of a Pharmaceutical Establishment | https://mohap.gov.ae/en/services/issue-a-certificate-of-compliance-with-the-good-practice-standards-of-a-pharmaceutical-establishment |
| Request to determine or modify the Narcotic drugs quotas of a Private Health or Pharmaceutical Institution | https://mohap.gov.ae/en/services/request-to-determine-or-modify-the-narcotic-drugs-quotas-of-a-private-health-or-pharmaceutical |
| Certificate of Release of Shipment of Pharmaceutical Products Derived from Biological Sources | https://mohap.gov.ae/en/services/certificate-of-release-of-shipment-of-pharmaceutical-products-derived-from-biological-sources |
| Request Treatment Abroad | https://mohap.gov.ae/en/services/request-treatment-abroad |
| Issue of quality report for medical product issued by Drug Quality Control Laboratory | https://mohap.gov.ae/en/services/issue-of-quality-report-for-medical-product-issued-by-drug-quality-control-laboratory |
| Licensing of Pharmacists | https://mohap.gov.ae/en/services/licensing-of-pharmacists |
| Issue an authorization to Export Narcotic Drugs | https://mohap.gov.ae/en/services/issue-of-permit-to-export-narcotic-drugs |
| Issue a Certificate of Compliance with good Manufacturing Practice Standards | https://mohap.gov.ae/en/services/issue-a-certificate-of-compliance-with-good-manufacturing-practice-standards |

| Title | Link |
|---|---|
| Evaluation of Pharmacovigilance Plan for Medical Products within UAE | https://mohap.gov.ae/en/services/evaluation-of-pharmacovigilance-plan-for-pharmaceutical-facility-and-its-subsidiaries |
| Approval of Medical Leaves and the leave of the patient's companion | https://mohap.gov.ae/en/services/approval-of-medical-leaves-and-reports |
| Submit Suggestions | https://mohap.gov.ae/en/services/submit-suggestions |
| Training of Pharmacists/Assistant Pharmacists in a Pharmaceutical Facility | https://mohap.gov.ae/en/services/training-of-pharmacistsassistant-pharmacists-in-a-pharmaceutical-facility |
| Issue of a 24-Hour Pharmacy Work Permit | https://mohap.gov.ae/en/services/issue-of-a-24-hour-pharmacy-work-permit |
| Submit Complaints and Feedback | https://mohap.gov.ae/en/services/submit-complaints-and-feedback |
| Submit Inquiries | https://mohap.gov.ae/en/services/submit-inquiries |
| Request for Health Statistics | https://mohap.gov.ae/en/services/applications-for-health-statistics-from-public-and-media |
| Request for support and coordination of health Research | https://mohap.gov.ae/en/services/request-for-support-and-coordination-of-medical-research |
| Mabrouk Ma Yak | https://mohap.gov.ae/en/services/mabrouk-ma-yak |

# APPENDIX-B

Table 10 - List of Test Questions and their corresponding ground truths

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| 1 | What is the purpose of the "Issue a Good Standing Certificate for a Health Professional" service? | This service allows medical staff from public and private sectors to apply for a certificate of good professional conduct. | MOHAP |
| 2 | What is the process for applying for a Good Standing Certificate? | 1. Log in to the MoHAP website or smart app using UAE PASS. 2. Submit the required documents based on the sector (public or private) and pay the applicable fee. 3. If the conditions are met, the certificate will be emailed and can also be downloaded from the website or app. | MOHAP |
| 3 | Who is not eligible for a Good Standing Certificate? | Trainees, visitors, individuals with only initial licenses, and those not licensed by the Ministry of Health and Prevention (MOHAP). | MOHAP |
| 4 | What are the general | The applicant must be licensed by | MOHAP |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| | requirements for obtaining a Good Standing Certificate? | MOHAP for more than 3 months, must not have been convicted of a medical offense, and the certificate is valid for 6 months. | |
| 5 | What are the specific document requirements for private sector medical staff? | Pharmacists and assistant pharmacists need an experience letter certified by the Department of Medicine in MOHAP. Holders of manual licenses need to specify the purpose of the application (i.e., obtaining a good standing certificate). | MOHAP |
| 6 | If a visiting foreign doctor wants to work in two UAE medical facilities and needs controlled drugs prescription books, what is the total cost they would need to pay for both licensing and obtaining the prescription books? | The total cost for a visiting doctor working in two facilities with controlled drugs prescription books would be AED 3,300. This comprises a visiting doctor license fee of AED 3,100 (including AED 100 application fee and AED 3,000 license fee) plus AED 200 for two controlled drugs prescription | |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| | | books (AED 100 each). | |
| 7 | A new private hospital with 75 beds wants to hire a visiting doctor from outside UAE for a year and set up a 24-hour pharmacy service. What would be the total initial licensing costs for all these services, including the hospital license? | For a new private hospital with 75 beds, hiring a visiting doctor and setting up 24-hour pharmacy service, the total initial licensing costs would amount to AED 39,300. This includes AED 30,000 for the hospital license (50-100 beds category), AED 100 application fee, AED 3,100 for the visiting doctor (AED 100 application + AED 3,000 license), and AED 6,100 for 24-hour pharmacy service (AED 100 application + AED 6,000 annual permit). | |
| 8 | A medical facility wants to register a conventional pharmaceutical product and later modify its dosage. What would be the total timeline and fees involved in both processes? | For registering and modifying a conventional pharmaceutical product, the total timeline would be 67 working days (45 days for initial registration plus 22 days for modification). The total fees would be | |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| | | AED 11,600, broken down as AED 10,600 for initial registration (AED 100 application + AED 7,000 registration + AED 3,500 analysis) and AED 1,000 for modification. | |
| 9 | If a pharmaceutical facility wants to change both its location and name in Dubai Healthcare City, what approvals would they need and what is the total cost for both changes? | For a pharmaceutical facility changing both location and name in Dubai Healthcare City, the total cost would be AED 2,200. This requires approvals from MOHAP, Dubai Healthcare City Authority, and Department of Economic Development. The cost breaks down as AED 1,100 for location change (AED 100 application + AED 1,000 fee) and AED 1,100 for name change (AED 100 application + AED 1,000 fee). | |
| 10 | For a medical facility that wants to add three new | Adding three new specialties and extending to 24- | |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| | specialties and extend to 24-hour service, what would be the complete workflow and total costs? | hour service would cost AED 24,200 total. The process takes approximately 2 working days for each specialty addition and 3 days for 24-hour service approval. The cost includes AED 18,000 for three specialties (AED 6,000 each), AED 100 application fee, and AED 6,100 for 24-hour service (AED 100 application + AED 6,000 annual fee). | |
| 11 | A new specialist clinic wants to start operations in UAE with initial clinic licensing, 24-hour pharmacy service, controlled drugs prescription book, and two visiting doctors from outside UAE. What would be the total initial cost? | For a new specialist clinic starting operations with initial licensing, 24-hour pharmacy service, controlled drugs prescription book, and two visiting doctors, the total initial cost would be AED 18,500. This includes AED 6,100 for initial clinic license (AED 100 + AED 6,000), AED 6,100 for 24-hour pharmacy (AED 100 + AED 6,000), AED 100 | |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| | | for controlled drugs book, and AED 6,200 for two visiting doctors (AED 200 applications + AED 6,000 licenses). | |
| 12 | A pharmaceutical facility needs to renew its registration of a conventional pharmaceutical product, change its location, and get approval for narcotic drugs quotas. What is the total timeline and fees? | A pharmaceutical facility renewing registration, changing location, and getting narcotic drugs quota approval would take 11-12 working days total and cost AED 4,800. This comprises AED 3,600 for product registration renewal, AED 1,100 for location change, and AED 100 for narcotic quotas. | |
| 13 | If a medical facility wants to add a new specialty and hire a visiting doctor for it, what documents would be needed and what is the maximum timeline? | Adding a new specialty and hiring a visiting doctor would take 7 working days total and cost AED 9,200. Required documents include a request letter and list of licensed staff for the specialty addition, plus job offer, license copy, | |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| | | facility plan, insurance, and undertaking letter for the visiting doctor. | |
| 14 | A hospital with 120 beds wants to renew its license, get approval for controlled drugs, and set up a 24-hour pharmacy service. What would be the annual cost and required approvals? | For a 120-bed hospital renewing its license, getting controlled drugs approval, and setting up 24-hour pharmacy service, the annual cost would be AED 46,300. This includes AED 40,100 for hospital renewal, AED 100 for controlled drugs, and AED 6,100 for 24-hour pharmacy. Required approvals include MOHAP license, Civil Defense, Economic Department, Drug Department, and waste management certification. | |
| 15 | What is the complete process and total cost for a medical facility that wants to change location, modify pharmaceutical product | The process of changing location, modifying pharmaceutical product registration data, and getting a good manufacturing practice certificate | |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| | registration data, and get a good manufacturing practice certificate? | would take 32-38 working days total. The base cost would be AED 7,100 (AED 4,100 for location change, AED 1,000 for registration modification, plus AED 2,000 per production line for the GMP certificate). | |
| 16 | If a new pharmaceutical facility wants to export a batch of locally manufactured narcotic drugs, what steps are necessary for compliance, and does the facility require any additional permits or documentation? | To export locally manufactured narcotic drugs, the facility must: 1. Initial Licensing: Be registered and licensed by MOHAP as a medical warehouse or manufacturer. 2. Narcotics Quota Approval: Submit Form F6 detailing quantities and types for an annual narcotics quota. 3. Export Permit: Apply for an export permit for narcotics with documentation of the approved quota. 4. Certificate of Pharmaceutical Product (CPP): Obtain this | |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
|  |  | certification from MOHAP confirming the product meets UAE standards for export. 5. Verification of Destination Compliance: Check if the destination country accepts UAE’s drug export certifications. |  |
| 17 | If a private clinic in the UAE wishes to change its name and simultaneously add a new specialty to its services, what are the procedural requirements and necessary documents for compliance with MOHAP regulations? | For a clinic to change its name and add a specialty, it must: 1. Change of Name: Apply through MOHAP with existing license, owner’s request letter, and updated trade license; the clinic must have been operational for at least six months. 2. Adding a Specialty: Submit a separate application specifying the new specialty and provide a list of relevant medical staff. 3. Verification: MOHAP will review both |  |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| | | applications for compliance. 4. Approval: Upon approval, the clinic receives updated licenses electronically. | |
| 18 | If I am appealing a decision regarding my medical licensing, how do I also apply for a renewal of my license to practice as a doctor? | To appeal a medical licensing decision, submit an appeal to the Office of the Minister of Health and Prevention within 15 days of receiving the decision. Simultaneously, to renew a license to practice as a doctor, log in to the MOHAP website, fill in the required information, attach necessary documents, and pay the renewal fee, ensuring compliance with continuous medical education requirements. | |
| 19 | What is the process for appealing a healthcare advertisement violation and subsequently updating the | To appeal a healthcare advertisement violation, submit an appeal to the relevant authority outlining the reasons for the | |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
|  | healthcare advertisement license for the same institution? | appeal. After addressing the appeal, if necessary, update the healthcare advertisement license by logging into the MOHAP system, completing the renewal form, and providing any additional documents or amendments required, alongside the application fee for the updated advertisement license. |  |
| 20 | What are the requirements to renew a healthcare advertisement license in a medical directory? How does the facility confirm the validity of its primary license before applying? | The facility must submit a renewal application for the healthcare advertisement license through the MOHAP website, attaching the necessary documentation and paying the application fee. To confirm the validity of its primary license, the facility must check its licensing status on the MOHAP system or contact the MOHAP support |  |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| | | to ensure all current licenses are valid and compliant with regulatory requirements. | |
| 21 | What is required to renew a pharmacist's license, and how does one also obtain an updated continuous education record for this renewal? | To renew a pharmacist's license, applicants must log in to the MOHAP website, provide the required documentation, and pay the renewal fee. They must also ensure that they have completed the necessary hours of continuous medical education (CME), which can be documented through participation in approved training programs or courses. A summary of CME hours completed must be attached to the renewal application. | |
| 22 | What is the process to register medical equipment? How can a company later apply for registration | To register medical equipment, the company must apply through the MOHAP website, submitting | |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| | renewal for continued sales in the UAE? | documentation such as the ISO certification and product details. The approval process typically takes about 45 working days. To renew the registration, the company must resubmit the original registration certificate along with any required updated documents and ensure compliance with current regulations, which can be done online through MOHAP's electronic service system. | |
| 23 | How can a complaint be lodged against a private health facility, and what are the follow-up requirements if the facility’s professional licenses are revoked due to the complaint? | Complaints against private health facilities can be submitted online via the MOHAP website or through customer service centers. If a facility’s professional licenses are revoked as a result of the complaint, follow-up may include reapplication for | |

| ID | Question | Ground Truths | Source |
|---|---|---|---|
| | | licenses, meeting additional compliance measures set by MOHAP, and undergoing inspections or assessments to ensure adherence to health regulations before being allowed to operate again. | |

# APPENDIX-C

Note that model names follow this Convention #eval-{Retriever}-{generator}-{chunksize}-@{k}

Table 11- Raw results of the evaluation of retriever models for both 528 and 1000 chunk-size databases.

| Model Name | Chunk Size | K | Ground Truth answer (semantic similarity) | IR hit rate | NDCG@k | Recall@k | latency | Total |
|---|---|---|---|---|---|---|---|---|
| eval_openai_openai_528_@5 | 528 | 5 | 0.700 | 0.696 | 1 | 0.255 | 6.322 | 0.742 |
| eval_gemini_openai_528_@5 | 528 | 5 | 0.713 | 0.652 | 1 | 0.249 | 5.558 | 0.736 |
| eval_cohere_openai_528_@5 | 528 | 5 | 0.726 | 0.652 | 1 | 0.192 | 5.765 | 0.732 |
| eval_cohere_openai_528_@3 | 528 | 3 | 0.723 | 0.565 | 1 | 0.156 | 4.976 | 0.704 |
| eval_openai_openai_528_@3 | 528 | 3 | 0.700 | 0.565 | 1 | 0.189 | 6.005 | 0.700 |
| eval_gemini_openai_528_@3 | 528 | 3 | 0.683 | 0.435 | 1 | 0.184 | 5.393 | 0.659 |
| eval_voyage_openai_528_@5 | 528 | 5 | 0.596 | 0.391 | 1 | 0.114 | 5.287 | 0.603 |
| eval_voyage_openai_528_@3 | 528 | 3 | 0.561 | 0.304 | 1 | 0.088 | 4.747 | 0.564 |
| **Model Name** | **Chunk size** | **K** | **Ground Truth answer (semantic similarity)** | **IR hit rate** | **NDCG@k** | **Recall@k** | **latency** | **Total** |
| eval_openai_openai_1000_@5 | 1000 | 5 | 0.743 | 0.478 | 1 | 0.131 | 9.268 | 0.687 |
| eval_gemini_openai_1000_@5 | 1000 | 5 | 0.755 | 0.435 | 1 | 0.118 | 7.052 | 0.678 |
| eval_cohere_openai_1000_@5 | 1000 | 5 | 0.787 | 0.348 | 1 | 0.098 | 7.175 | 0.666 |
| eval_gemini_openai_1000_@3 | 1000 | 3 | 0.735 | 0.391 | 1 | 0.097 | 6.092 | 0.656 |
| eval_voyage_openai_1000_@5 | 1000 | 5 | 0.635 | 0.522 | 1 | 0.117 | 6.105 | 0.652 |
| eval_openai_openai_1000_@3 | 1000 | 3 | 0.735 | 0.348 | 1 | 0.082 | 7.439 | 0.643 |
| eval_cohere_openai_1000_@3 | 1000 | 3 | 0.691 | 0.217 | 1 | 0.059 | 7.389 | 0.590 |
| eval_voyage_openai_1000_@3 | 1000 | 3 | 0.565 | 0.261 | 1 | 0.067 | 5.823 | 0.551 |

# APPENDIX-D

Multihop and Yes-No Questions

Table 12 - Raw results of the evaluation of generator model in Naive RAG for Multihop and Yes-No Questions

| Model Name | Reco. | Average Latency | Answer Relevance | Groundedness | Context Relevance | Weighted Average |
|---|---|---|---|---|---|---|
| Claude 3.5 Sonnet | 24 | 10.892 | 1.000 | 0.907 | 0.531 | 0.869 |
| Openai 4o | 24 | 5.593 | 0.931 | 0.958 | 0.542 | 0.864 |
| Claude_3.5-Haiku | 24 | 9.692 | 0.958 | 0.923 | 0.547 | 0.862 |
| openai 4o mini | 24 | 6.106 | 1.000 | 0.879 | 0.547 | 0.861 |
| gemini_1.5_flash | 24 | 3.029 | 0.903 | 0.941 | 0.536 | 0.845 |
| command_r_plus | 24 | 4.992 | 0.944 | 0.828 | 0.544 | 0.818 |
| gemini_1.5_pro | 24 | 5.448 | 0.681 | 0.968 | 0.531 | 0.766 |

Single Hop Direct Questions

Table 13 - Raw results of the evaluation of generator model in Naive RAG for Single Hop Questions

| Model Name | Reco. | Average Latency | Answer Relevance | Groundedness | Context Relevance | Weighted Average |
|---|---|---|---|---|---|---|
| gemini 1.5 pro | 45 | 5.374 | 1 | 0.975 | 0.735 | 0.937 |
| gemini 1.5 flash | 45 | 2.661 | 0.993 | 0.963 | 0.738 | 0.930 |
| Openai 4o | 45 | 5.094 | 1 | 0.952 | 0.739 | 0.929 |
| Claude 3.5 Haiku | 45 | 9.777 | 1 | 0.942 | 0.745 | 0.926 |
| Claude 3.5 sonnet | 45 | 9.782 | 0.993 | 0.945 | 0.73 | 0.921 |
| Openai 4o mini | 45 | 4.785 | 1 | 0.911 | 0.732 | 0.911 |
| Command r plus | 45 | 5.224 | 1 | 0.862 | 0.735 | 0.892 |

# APPENDIX-5

Hybrid RAG:

Table 14 - Raw results of the evaluation of generator model in Hybrid RAG

| App Version | alpha | Records | Average Latency | Answer Relevance | Groundedness | Context Relevance | Weighted Average |
|---|---|---|---|---|---|---|---|
| roberta-4o-large_3-500-alpha_0.25-02 | 0.25 | 23 | 3.718 | 0.71 | 0.649 | 0.246 | 0.593 |
| roberta-4o-large_3-500-alpha_0.5 | 0.5 | 23 | 3.881 | 0.725 | 0.726 | 0.249 | 0.630 |
| roberta-4o-large_3-500-alpha_0.75 | 0.75 | 23 | 4.241 | 0.783 | 0.697 | 0.267 | 0.645 |
| roberta-4o-large_3-500-alpha_0.95 | 0.95 | 23 | 3.65 | 0.71 | 0.788 | 0.33 | 0.665 |
| roberta-4o-large_3-500-alpha_1.0 | 1 | 23 | 4.208 | 0.971 | 0.941 | 0.675 | 0.900 |

# REFERENCES

Wu, M., & Cao, S. (2024). LLM-Augmented Retrieval: Enhancing retrieval models through language models and Doc-Level embedding. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2404.05825

Wang, F., Wan, X., Sun, R., Chen, J., & Arık, S. Ö. (2024). Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2410.07176

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2023). SELF-RAG: Learning to retrieve, generate, and critique through self-reflection. arXiv. https://arxiv.org/abs/2310.11511

Izacard, G., & Grave, E. (2021). Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. arXiv preprint arXiv:2007.01282. https://arxiv.org/abs/2007.01282

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. arXiv preprint arXiv:2104.08663. https://arxiv.org/abs/2104.08663

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense Passage Retrieval for Open-Domain Question Answering. arXiv preprint arXiv:2004.04906. https://arxiv.org/abs/2004.04906

groundtruth - □ TruLens. (n.d.). https://www.trulens.org/reference/trulens/feedback/groundtruth/#trulens.feedback.groundtruth.GroundTruthAgreement.ndcg_at_k

Gao, L., Ma, X., Lin, J., & Callan, J. (2023). Precise Zero-Shot Dense Retrieval without Relevance Labels. Proceedings of the 61st Annual Meeting of the Association for Computational

Linguistics (Volume 1: Long Papers), 1762–1777. https://doi.org/10.18653/v1/2023.acl-long.99

Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2022). MuSiQue: Multihop questions via single-hop question composition. Transactions of the Association for Computational Linguistics, 10, 1 633–648

deepset. (n.d.). Query expansion: Techniques to improve retrieval. Retrieved from https://haystack.deepset.ai/blog/query-expansion

OpenAI. (2024, January 25). New embedding models and API updates. OpenAI. Retrieved from https://openai.com/index/new-embedding-models-and-api-updates/

Cohere. (2023, November 2). Introducing Embed v3. Cohere. Retrieved from https://cohere.com/blog/introducing-embed-v3

Google AI. (2024, November 19). Embeddings in the Gemini API. Google AI for Developers. Retrieved from https://ai.google.dev/gemini-api/docs/embeddings

Voyage AI. (2024, June 10). voyage-multilingual-2: Multilingual Embedding Model. Voyage AI Blog. Retrieved from https://blog.voyageai.com/2024/06/10/voyage-multilingual-2-multilingual-embedding-model/