For this assignment, the task I chose is task 4 which doing the query reformulation. In order to expand the efficiency and performance of the query during the query searching process, I decided to use the IR system. Since we know that the query will get more score and rank higher if it is more relevant to the visiting documents, I decided to let the IR system rephrase a query by using the prompting techniques to modify it and doing the text processing to improve the recall metric of the IR system, ensuring it efficiently retrieves all relevant documents corresponding to a user's search intent. In order to get the modified version of the query, I put the original query without doing any modification in the ChatGPT Web Application and let it use the prompting techniques like zero-shot prompting to get the output which is the more comprehensive query designed to retrieve a broader range of relevant documents from the IR system. The effectiveness of this query reformulation will be evaluated using standard IR metrics like recall, precision, or map, by contrasting the system's performance with the original and revised queries.

The evaluation metric I decided to use to measure performance is the Mean Average Precision (MAP) metric. The reason why I choose the MAP is that it can account for both the accuracy and rank of retrieved relevant documents across multiple queries which can comprehensively measure the IR system's effectiveness and takes the average of all the Average Precision score across all queries to reflect both the ability to retrieve relevant documents (recall) and to rank them highly (precision). Since I decided to automate evaluation with the GPT, the evaluation prompt techniques I use are zero-shot prompt, few-shot prompt, and chain-of-thought prompt and I collect the results of the judgements for both GPT and me on the contingency table below.

| | # My relevant | # My not relevant |
|---|---|---|
| # GPT's relevant | 51 | 9 |
| # GPT's not relevant | 6 | 234 |

According to the contingency table, we can have:

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)}$$
$$P(A) = \frac{51 + 234}{51 + 9 + 6 + 234} = \frac{285}{300} = 0.95$$
$$P(E) = P(Relevant)^2 + P(Non\ relevant)^2$$
$$P(Relevant) = \frac{51 + 6 + 51 + 9}{300 \times 2} = \frac{117}{600} = 0.195$$
$$P(Non\ relevant) = \frac{6 + 234 + 9 + 234}{300 \times 2} = \frac{483}{600} = 0.805$$
$$P(E) = P(Relevant)^2 + P(Non\ relevant)^2 = 0.195^2 + 0.805^2 = 0.68605$$
$$Kappa = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.95 - 0.68605}{1 - 0.68605} = 0.841$$

The Kappa value we got is 0.841 which falls within the range of almost perfect agreement which suggests that the evaluations provided by me and GPT are in very strong concordance with one another after accounting for the possibility of chance agreement which implies that the method

or system being used for classification or judgement is reliable and would be expected to produce similar results upon repeated applications.

The task I chose is task 4 which is the task of reformulation and augmentation is a critical component in the Decision Support System as it directly influences the system's ability to provide relevant information, which is vital for making informed decisions. For example, in healthcare, a well-refined query can help medical professionals retrieve the most current and comprehensive research, leading to better patient care strategies. In the legal domain, precise and expanded queries enable lawyers to find pertinent case laws and precedents, ensuring more robust legal arguments and strategies. By improving recall, this task ensures that decision-makers have access to a wider array of information, minimizing the risk of overlooking critical data. Hence, task 4 is not only about enhancing search efficiency; it is about empowering users with the right information at the right time, which is the cornerstone of effective decision-making.

In order to have more chances to get the more relevant documents to which the queries correspond and get better search results, I used zero-shot prompting to modify our original queries first. For example, if the original query is “Renewable energy”, it will be modified by the prompting technique and turn to “Latest advancements in solar and wind energy technologies” and this new version of the query will be used as the input. After getting the modification of the query, we put this query into our IR system. In our IR system, we will apply the text processing method to modify the input in order to let the input get into a consistent format to help users effectively match queries with the most pertinent documents in the database. The text processing methods I applied in the IR system are tokenization, lowercasing, removing stop words, and lemmatization. After applying the text processing method, we will get the final desired output. For example, if our input is “Latest advancements in solar and wind energy technologies”, our desired output is “latest advancement solar wind energy technology”. Finally, the IR system will use these desired outputs to determine the extent of the query relevant to the documents and use the evaluation metrics to measure their performances. I write the table of the inputs and desired outputs for the test examples of my task below.

| Input | Desired output |
|---|---|
| Latest advancements in solar and wind energy technologies | latest advancement solar wind energy technology |
| Impact of artificial intelligence on patient diagnosis accuracy | impact artificial intelligence patient diagnosis accuracy |
| Consequences of Arctic ice melt on global sea levels | consequence arctic ice melt global sea level |
| Applications of quantum computing in cryptography | application quantum computing cryptography |
| Effectiveness of crop rotation in sustainable farming practices | effectiveness crop rotation sustainable farming practice |

Since we want all queries can get better search results, I decided to use three prompt styles which are zero-shot prompting, few-shot prompting, and chain-of-thought prompting. Firstly, the basic zero-shot prompt directly instructs the system to reformulate the query without additional context or examples. It is a minimal approach to see how well the system can interpret and improve the query based on its pre-trained knowledge alone. The template for implementing the basic zero-shot prompting is below.

| Zero-shot prompting template |
| --- |
| Query: [Input Query]<br>Reformulate this query to enhance search effectiveness. |

For example, if we use “Renewable energy” as the input query, we will get the modified query “Latest advancements in solar and wind energy technologies” after reformulating. Secondly, the few-shot prompt includes examples of original and reformulated queries before presenting the input query. These examples serve as context, helping the system understand the desired pattern of reformulation. The template for implementing the basic few-shot prompting is below.

| Few-shot prompting template |
| --- |
| Example 1:<br>Original Query: "Renewable energy"<br>Reformulated: "Comparative analysis of efficiency in renewable energy sources"<br><br>Example 2:<br>Original Query: "AI in healthcare"<br>Reformulated: "Review of AI applications in personalized medicine"<br><br>Query: [Input Query]<br>Reformulate this query based on the examples given. |

For example, if we use “Climate change in Arctic” as the input query, we will get the modified query “Effects of climate change on Arctic marine biodiversity” after reformulating. Thirdly, the chain-of-thought prompt guides the system through a thought process for reformulating the query. It encourages considering broader or more specific aspects related to the main topics of the query, leading to a more thoughtful and potentially effective reformulation. The template for implementing the basic chain-of-thought prompting is below.

| Chain-of-thought prompting template |
| --- |
| Query: [Input Query]<br>To improve the recall of this query, consider the main topics and related key terms. Think about broader or more specific aspects that could be included in a more effective search, such as “[Input Information]”.<br><br>Reformulated Query: |

For example, if we use “Quantum computing” as the input query and “quantum computing use cases in cryptography and data security” as the input information, we will get the modified query “Potential of quantum computing in enhancing encryption and data security” after reformulating. I hope that these three prompt styles I mentioned above will get better search results for all queries and get better evaluation scores.

After implementing the prompt techniques to modify the queries, we use the modified queries as the inputs and put these inputs in our IR system to determine the extent of relevance to the documents and use the MAP metric to test their performance. The three table below shows the performance of the queries by using the three different prompt techniques.

| MAP score by using the zero-shot prompting | |
|---|---|
| Query | MAP score |
| Latest advancements in solar and wind energy technologies | 0 |
| Impact of artificial intelligence on patient diagnosis accuracy | 0 |
| Consequences of Arctic ice melt on global sea levels | 0 |
| Applications of quantum computing in cryptography | 1 |
| Effectiveness of crop rotation in sustainable farming practices | 0 |
| Emerging trends in cybersecurity threats in the financial sector | 0 |
| COVID-19's long-term effects on small business sustainability | 0 |
| Influence of social media on adolescent self-esteem | 0 |
| Comparative analysis of battery life in 2023 electric vehicle models | 0 |
| Conservation strategies for endangered species in the Amazon rainforest | 0 |
| Overall MAP score: 0.1 | |

| MAP score by using the few-shot prompting | |
|---|---|
| Query | MAP score |
| Comparative analysis of efficiency in renewable energy sources | 0 |
| Review of AI applications in personalized medicine | 0 |

| | |
|---|---|
| Effects of climate change on Arctic marine biodiversity | 0 |
| Current challenges in commercial application of quantum computing | 0 |
| Case studies on reducing water usage in sustainable agriculture | 0 |
| Analysis of cybersecurity threat patterns in cloud computing | 0 |
| Studies on the economic recovery post-COVID-19 pandemic | 0 |
| Social media's effects on communication skills in young adults | 0 |
| Impact of electric vehicles on urban air quality improvement | 0 |
| Role of habitat conservation in protecting endangered species | 1 |
| Overall MAP score: 0.1 | |

| MAP score by using the chain-of-thought prompting | |
|---|---|
| Query | MAP score |
| Evaluation of wind energy vs. solar energy in residential power generation | 1 |
| The role of AI in improving diagnostic accuracy for early-stage diseases | 0.2 |
| Study on the rate of ice melt in the Arctic and its impact on global sea levels | 0.2 |
| Potential of quantum computing in enhancing encryption and data security | 1 |
| Innovations in sustainable agriculture for arid climates | 1 |
| Emerging cybersecurity threats in IoT devices and preventative strategies | 0.5 |
| Analysis of COVID-19's impact on the global supply chain in the tech industry | 1 |
| Investigation into the effects of social media on attention span and memory retention | 0 |
| Comparison of electric vehicle adoption rates in urban versus rural areas | 1 |
| Conservation efforts for endangered marine species in the Great Barrier Reef | 1 |
| Overall MAP score: 0.69 | |

From the results tables above, firstly, for the zero-shot prompting, the overall MAP score for zero-shot prompting is 0.1, indicating that for most queries, the system did not retrieve relevant documents at the top of the rankings. This suggests that zero-shot prompting, without additional context or examples, may not have provided enough guidance for the system to understand and retrieve the most relevant documents for each query. And then, for the few-shot prompting, the MAP score remains 0.1 for few-shot prompting as well, which is identical to the zero-shot prompting. This could imply that the few examples given were not sufficiently informative or that the system could not generalize well from the provided examples to the test queries. It is also possible that the few-shot examples were not closely related to the test queries, leading to no significant improvement in performance. However, for the chain-of-thought prompting, there is a notable increase in the overall MAP score to 0.69. This indicates a significant improvement in retrieving relevant documents. This method seems to help the system by providing a reasoning framework that aligns better with the information retrieval task, leading to more accurate results. Therefore, it means that comparing with the zero-shot and few-shot prompting styles, the chain-of-thought prompting style provided the best performance in terms of MAP score, suggesting that guiding the system through a reasoning process can significantly enhance the retrieval of relevant documents.

After the overall discussion, although the chain-of-thought prompting style has the best performance, some of the queries from the examples still do not perform really well by using this style. The two tables below show three examples in which LLM did well and three examples in which LLM did poorly, respectively.

| Examples of LLM did well by using the chain-of-thought prompting | |
|---|---|
| Query | MAP score |
| Potential of quantum computing in enhancing encryption and data security | 1 |
| Analysis of COVID-19's impact on the global supply chain in the tech industry | 1 |
| Conservation efforts for endangered marine species in the Great Barrier Reef | 1 |

| Examples of LLM did poorly by using the chain-of-thought prompting | |
|---|---|
| Query | MAP score |
| The role of AI in improving diagnostic accuracy for early-stage diseases | 0.2 |
| Study on the rate of ice melt in the Arctic and its impact on global sea levels | 0.2 |
| Investigation into the effects of social media on attention span and memory retention | 0 |

According to the two tables above, I think that the reason why the LLM performed well for the queries in the first table above is that these queries are well-defined and refer to specific, narrow topics that align well with documented and researched subjects. The specificity of the

queries allows the LLM to latch onto key terms that are likely to be present in relevant documents, leading to high precision in retrieval. Comparing with the first table above, the reason why the LLM did poorly for the queries in the second table above is that these queries may be too broad or require a deep understanding of nuanced relationships within the topics. For instance, the effects of social media on cognitive functions can be a diffuse topic with many conflicting findings, making it harder for the LLM to identify clear, relevant documents. The lower MAP scores indicate that relevant documents either were not retrieved or ranked lower in the search results, suggesting that the LLM had difficulty in understanding or associating the queries with the appropriate documents. Therefore, it means that the successful examples seem to benefit from a focused scope and clear terminology that closely matches the content of relevant documents. In contrast, the unsuccessful examples may involve complex, multi-faceted topics or topics that are too broad, resulting in the retrieval of less relevant information. The chain-of-thought prompting may have assisted the LLM in the successful examples by guiding it through a structured reasoning process, which could have helped in forming better query reformulations for precise information retrieval.

Although LLM seems to be very efficient and useful, it still has some limitations. When deploying an LLM for query reformulation, one issue that could arise is the LLM's potential misunderstanding of the context or the specific domain of the query. This can lead to reformulated queries that are off-topic or irrelevant, thus diminishing the quality of search results. To handle this in practice, one could incorporate domain-specific fine-tuning or provide the LLM with domain-specific knowledge bases, ensuring it has the necessary context to understand and reformulate queries accurately. Another issue could be the inconsistency in the quality of query reformulations. Since zero-shot and few-shot prompts may not always provide sufficient guidance, and chain-of-thought prompting depends on the LLM's ability to generate a logical reasoning path, the LLM might sometimes produce reformulations that do not improve or even degrade search performance. To mitigate this, a feedback loop could be established where the system learns from successful and unsuccessful reformulations over time, possibly augmented by human-in-the-loop interventions to guide the LLM towards better performance. However, although LLM still has some limitations, it is still a very powerful technique to improve the performance of the Decision Support System.

Appendix

| Input | Desired output |
|---|---|
| Emerging trends in cybersecurity threats in the financial sector | emerging trend cybersecurity threat financial sector |
| COVID-19's long-term effects on small business sustainability | effect small business sustainability |
| Influence of social media on adolescent self-esteem | influence social medium adolescent |
| Comparative analysis of battery life in 2023 electric vehicle models | comparative analysis battery life electric vehicle model |
| Conservation strategies for endangered species in the Amazon rainforest | conservation strategy endangered specie amazon rainforest |

Table 1: 5 test examples for question 2