Special Topics in Data Science
Prompt Engineering: Art and Science for Enhancing QA Interactions with Large Language Model
Final Project Report

December 2024

Students Cecilia Nymberg, Yoomin Kim, Zakaria Chahboune
Group CYZ
Professor LI WEN SYAN

1 Introduction

This project aims at accurately answering financial questions, which is critical for effective decision-making. This study evaluates a query-answering model designed to provide precise responses to questions based on a structured financial dataset. The dataset includes historical company data and key metrics. By analyzing its performance on simple and complex queries, we aim to identify ways to improve accuracy and reliability.

2 Designing the Planning Agent

The task of the planning agent is to analyze the query and decide what steps to take to answer the query. To do this, it has a think-tank for guidance.

2.1 Query Decomposition

However, before turning to the think-tank, the planning agent decomposes the query into subqueries. Some queries consists of several queries (multiqueries) and these needs to be broken down before relevant documents can be retrieved. Inspiration for implementing the query decomposer was found here https://haystack.deepset.ai/blog/query-decomposition. The query decomposer consists of an LLM specially instructed for query decomposition. The LLM was prompted using few-shot examples. It also uses a Questions schema for structured output. After decomposing the query, the subqueries are sent back to the planning Agent.

The prompt for the query decomposer can be seen in the lstlisting below.

```
You are a helpful assistant that prepares queries that will be sent to a
search component. Sometimes, these queries are very complex. Your job is
to simplify complex queries into multiple queries that can be answered in
isolation to each other.

Question: {query}

If the query is simple, then keep it as it is.

Examples
1. Query: Did Apple or American Tower Corp make more money last year?
   Decomposed Questions:
   [Question(question='How much profit did Apple make last year?', answer=None),
    Question(question='How much profit did American Tower Corp make last year?',
    answer=None)]

2. Query: What is the difference of Operating Profit Margin between Air Products
   and Chemicals in 2016 and printing papers business of International Paper
   Company in 2001?
   Decomposed Questions:
   [Question(question='What was the Operating Profit Margin Air Products and
   Chemicals in 2016?', answer=None)],
```

```
    [Question(question='What was the Operating Profit Margin printing papers
    business in 2001?', answer=None)]

3. Query: {{question}}
    Decomposed Questions:
```

Listing 1: Query Simplification Prompt

Example usage:

```
decomposer_chain.invoke({
    "query": "What is the difference of Union Pacific Corp's revenue from 2013 to
        2016 divided "
             "by the sum of Operating Profit Margin between Air Products and
                 Chemicals in 2015 "
             "and printing papers business of International Paper Company in 2004?"
})
Output:
Questions(
    questions=[
        Question(question="What was the revenue of Union Pacific Corp in 2013?"),
        Question(question='What was the revenue of Union Pacific Corp in 2016?'),
        Question(question='What was the Operating Profit Margin of Air Products and
            Chemicals in 2015?'),
        Question(question='What was the Operating Profit Margin of the printing
            papers business of International Paper Company in 2004?')
    ]
)
```

Listing 2: Query Decomposition Example

2.1.1 Tools for think tank

The implemented tools in the think-tank are the following:
Domain specific:

- `calculate_operating_profit_margin`
- `calculate_percentage_change`
- `calculate_ratio`
- `calculate_difference`
- `rewrite2fiscal`
- `calculate_eps`
- `calculate_cashflowfromoperations`

Other:

- `retrieve_factual_data`
- `classify_query`
- `advanced_retrieval`

The `classify_query` tool categorizes queries into low, medium, or high difficulty. This makes it possible to use query routing. Initially, it over-classified most queries as high difficulty. To solve this, database retrieval capabilities was given to the classifier. This was it could more easily determine the right difficulty level.

The classifyer was prompted as can be seen below:

```
system_query = """
Your task is to classify the financial query as low_level, medium_level, or
    high_level.
Query: {query}.
```

```

Use the following strategy:
1. Check if the query is a low level query:
   - Use retrieve_factual_data to see if the necessary context can be retrieved
     using simple RAG.
   - If the retrieved information can answer the query, output low_level.

2. Check if the query is a medium level query:
   - Apply relevant tools: calculate_eps, calculate_cashflowfromoperations,
     operating_profit_margin, and percentage_change.
   - Then use retrieve_factual_data to retrieve input for the tools or other context.
   - If so, output medium_level.

3. Check if the query is a high level query:
   - Use advanced RAG to retrieve the necessary context.
   - If context cannot be retrieved using retrieve_factual_data or tools,
     output high_level.

Important: Only output the difficulty level without any additional information.
Output: {{difficulty_level}}
"""
```

Listing 3: Query Classification Strategy

A classification example can be seen below:

```
classify_query("What is the difference of Operating Profit Margin between Air
    Products
and Chemicals in 2016 and printing papers business of International Paper Company in
    2006?")

user_query Question: What is the difference of Operating Profit Margin between Air
    Products
and Chemicals in 2016 and printing papers business of International Paper Company in
    2006?
ticker: APD
fy: 2017
Retrieved documents:
[Document(metadata={'company': 'IP', 'context_type': 'post_text', 'fiscal': 2006},
page_content='. operating profits in 2006 were up 31% (31%) compared with 2005 and 6%
     (6%)
compared with 2004. this ...')]
'medium_level'
```

Listing 4: Query Classification Example

2.2 Chain-of-though

Chain of Thought (CoT) was implemented to instruct the planning agent on how to structure the task of answering the query. First, it was directed to decompose the query into subqueries. Then, it classified each subquery by difficulty and used the appropriate tools to gather context and formulate answers. Finally, it combined these answers to construct a complete and concise response to the original query. The prompt using CoT is presented in the lstlisting below.

```
You are a professional agent that can use tools to answer a financial question.
To answer the questions, you can retrieve documents. The documents consist of
text and tables. The tables have row names and values; look for relevant row
names to answer the query.

Use the following strategy to answer the question:
1. Decompose the query using the decompose_query tool.
2. For each subquery in the decomposed query:
   1. Classify the subquery using classify_query.
      If the subquery is low_level:
          Use retrieve_factual_data to get context and answer the subquery.
```

```
      Elif the subquery is medium_level:
          Use relevant tools. Use retrieve_factual_data to get the necessary
          context. Then answer the subquery.
      Elif the subquery is high_level:
          Use relevant tools if needed. Use advanced_retrieval to get the
          necessary context. Then answer the subquery.

3. Use the answers to the subquestions to answer the original question.
   Provide a concise but complete answer.
If you struggle with finding information, you may take help from the
advanced_retrieval even for low or medium level questions.
```

Listing 5: System Query Prompt

2.3 Retrieval Agent

The retreieval agent was prompted as can be seen below. The prompt uses both exemplairs and Chain-of-Though. It is also prompted to use query expansion by using synonyms to create more queries.

```
system_query = """
Your task is to retrieve important contextual documents to answer the financial query
    .
The queries are a bit tricky, so you have to use smart techniques to find the
    relevant documents.

Query: {query}.

Use the following strategy:
1. Filter out keywords in the query and use synonyms.
   Example:
   Query: Air Products current ratio 2024
   Keywords:
     - current ratio
     - liquidity ratio
     - cash flow ratio
     - working capital ratio
     - liquidity
   Also, try to keep the query short when retrieving documents.

2. Shorten complex queries.
   Example:
   Query: What was the Operating Profit Margin of the printing papers business
          of International Paper Company in 2005?
   Shortened query: Profit printing International Paper Company 2005

3. Conduct the search using the search tools: hybrid_search or retrieve_factual_data.
   - If you find a relevant unit in the query or can guess it, search using that.
   Example:
   Query: What was the Operating Profit Margin of the printing papers business
          of International Paper Company in 2005?
   Unit query: Paper Company 2005 million dollars $

4. Try different approaches until you retrieve the needed context.
   - You can also try using the previous or next year if relevant.

5. Once you find the context, retrieve the relevant documents.

For your help, you can call tools.
Output: {{Relevant context}}
"""
```

Listing 6: Retrieval Agent Prompt

Additionally, the retriver can use the following tools.

- rewrite2fiscal,
- hybrid_search
- retrieve_factual_data

The retrieve_factual_data function uses semantic search but filters the search on different factors. The function retrieves documents using semantic similarity search but filters the search in different ways. First, it retrieves three documents, filtering based on the company ticker (`ticker`) and fiscal year (`fy`). If a `unit` is specified, it appends the unit to the query and performs another search, filtering specifically for tables associated with the specified company. Similarly, if a `keyword` is provided, it searches for tables filtered by the company, column names, and context type. Another search is conducted to retrieve two table documents based on the company, fiscal year, and context type. The results from these searches are combined, prioritized, and optionally enriched with unit and keyword-based results. Finally, the retrieved documents are concatenated into a single text for further processing.

The `hybrid_search` function retrieves context-related documents by combining different retrieval methods. It accepts a query string as input and returns the retrieved documents. It creates two retrievers: a BM25 retriever using `BM25Retriever.from_documents(documents)` and a vector-based retriever using `docsearch.as_retriever()`. These retrievers are combined into an `EnsembleRetriever`, which assigns equal weights (`[0.5, 0.5]`) to both retrieval methods. The ensemble retriever processes the query to fetch relevant documents, prints the results, and returns them.

The reason why the hybrid search can be benefical is beacuse it involves an alternative way to search. The BM25Retreiver searches using exact keyword match instead of similarity search. This way, it can find other results than the VectorDB retriever. The implementation of the hybrid_search can be seen below.

```python
@tool
def hybrid_search(query: str) -> str:
    """
    Uses hybrid search to retrieve context related documents.

    Args:
    query (str): The question to be answered.

    Returns:
    str: The retrieved documents.
    """
    print("using hybrid search")
    print("query", query)
    bm25_retriever = BM25Retriever.from_documents(documents)
    vector_retriever = docsearch.as_retriever()

    retrievers = [bm25_retriever, vector_retriever]
    ensemble_retriever = EnsembleRetriever(
        retrievers=retrievers,
        weights=[0.5, 0.5]
    )
    result = ensemble_retriever.get_relevant_documents(query)
    return result
```

Listing 7: Hybrid Search Function

The retriever Agent is also told to only retrieve the relevant documents, to make it easier for the generating agent.

3 Other Approaches

While working on this project we tried many methods and techniques, and only kept the most efficient ones in the end. As a matter of facts, we tried using an LLM agent for classifying the queries into levels of complexity. However this method only added more complexity to the model and its implementation, and was in the end less efficient that the Chain-of-thought finally implemented.

Moreover, we used a pipeline of two LLM agents to handle temporal variables. The goal was to re-write the queries that contain relative time like 'last year', or '5 years ago', into new queries with absolute time that the LLM would understand. The role of the first agent was finding the elements of relative time, and then deduce

the 'absolute time', representing the actual timing targeted by the query. The second agent had to use the query, the relative time, and the absolute time to generate a new query, as a pre-processing before the sub-division and retrieval tasks. However this unfortunately only introduced a layer of randomness to the model. In facts, it would always have a hard time creating the new query without flaws, even if the absolute date was successfully retrieved in the first step. Using this method lowered the accuracy to 23/50, and produced a random error after 13 minutes in another attempt. A deeper research on handling the temporality would have certainly enhanced the final accuracy.

4 Results and Analysis

Our model reaches an accuracy of 54%, as can be seen in table 1. This result is rather correct considering the complexity of the database and of some the questions. In most cases, the model correctly targets the good year, company, and information thanks to the multiple layers of processing, and using various tools. However some errors occur due to inaccuracies in the process. As a matter of facts, we use various LLMs one after the other, which increases the risk of propagating an error in the chain. Moreover du to the unpredictability of LLMs, it is almost impossible to obtain exactly the expected result every time.

Table 1: Classification Accuracy

Metric	Value
Accuracy	54%
Correctly Classified Questions	27/50

```
If you need further assistance or have other questions, please let me know!
Accuracy: 27/50
```

Figure 1: Accuracy

Correct Questions

A total of 26 questions were found to be correct. These questions were characterized by accurate input data, where the requested information clearly existed in the database. They typically involved simple calculations or straightforward data retrieval tasks, such as covering a single year or specific financial items. For example, the question "What is the aggregate rent expense of American Tower Corp in 2014?" was correctly answered as "$655.0 million dollars." Similarly, "What is the long-term component of BlackRock at 12/31/2011?" was accurately answered as "593356 million dollars." This high accuracy is attributed to the system's ability to handle focused queries involving specific years and financial items, leveraging reliable search and calculation tools.

Wrong Answers

A total of 24 questions were found to be incorrect, primarily due to three main issues. First, some questions required data that was not available in the database. For instance, the question "What is Entergy Corp's 2008 total value, in millions of dollars, of issuable long-term securities?" could not be answered because the relevant information was missing. Second, questions involving complex calculations, such as multi-step processes or ratios, often led to errors. An example is "What is the difference of Operating Profit Margin between Air Products and Chemicals in 2015 and printing papers business of International Paper Company in 2006?" where input inconsistencies caused inaccuracies. Finally, unclear questions with convoluted structures or ambiguous keywords posed significant challenges. For instance, the overly complex query "What is the Goldman Sachs Group's Total Assets change from 2013 to 2017 divided by the sum of Operating Profit Margin between Air Products and Chemicals in 2016 and printing papers business of International Paper Company in 2004?" caused confusion during both the search and calculation phases.

Category	Number of Correct Answers	Number of Incorrect Answers	Ratio (%)
Simple search questions	16	2	88.9
Simple calculation questions	7	5	58.3
Complex calculation questions	3	17	15.0

Table 2: Performance of the system across different question categories.

Conclusion

We implemented a complex model using LLM agents and various tools and prompt techniques to answer precise questions of finance. We used a pipeline to process the query, and extracted the relevant information using tools. Then we retrieved the information in the database and eventually processed it again depending of the nature of the query, and using agent tools again. We reached a correct accuracy that could be enhanced using even more advanced and complex methods, by creating new pipelines and tools for each possible case, or by better handling the dimension of time in the queries.

5 Appendix

Category	Content
Wrong	Question 11: From the perspective of 5 years ago, what percentage of total minimum lease payments of Dish Network are due in 2015?
Wrong	Question 40: What is the difference of current ratio between American Tower Corp in 2012 and DISH Network Corporation in a year before that?
Wrong	Question 14: Based on the cash dividends paid of the year, how many common stock shares were outstanding in 2007 Snap-on?
Wrong	Question 15: What is the Entergy Corp's 2008 total value, in millions of dollars, of issuable long-term securities?
Wrong	Question 16: For the capital framework of Goldman Sachs Group 7 years ago, what percent of the minimum supplementary leverage ratio consisted of a buffer?
Wrong	Question 19: In 2014 Global Payments, what is the total value of securities approved by security holders but not yet issued (in millions)?
Wrong	Question 25: What is the International Paper Company's Operating Profit Margin of printing papers business in 2006?
Wrong	Question 35: What is the difference of Operating Profit Margin between Air Products and Chemicals in 2015 and printing papers business of International Paper Company in 2006?
Wrong	Question 36: What is the average of Operating Profit Margin between Air Products and Chemicals in 2015 and printing papers business of International Paper Company in 2005?
Wrong	Question 37: What is the sum of Operating Profit Margin between Air Products and Chemicals in 2015 and printing papers business of International Paper Company in 2004?
Wrong	Question 41: What is the Goldman Sachs Group's Total Assets change from 2014 to 2016 multiplied by the average of current ratio between American Tower Corp in 2012 and DISH Network Corporation in 2011?
Wrong	Question 33: What is the sum of Operating Profit Margin between Air Products and Chemicals in 2016 and International Paper Company in 2004?
Wrong	Question 42: What is the Goldman Sachs Group's Total Assets change from 2013 to 2017 multiplied by the average of current ratio between American Tower Corp in 2012 and DISH Network Corporation in 2011?
Wrong	Question 43: What is the sum of Goldman Sachs Group's Total Assets from 2011 to 2014 divided by the average of Operating Profit Margin between Air Products and Chemicals in 2016 and printing papers business of International Paper Company in 2005?

Category	Content
Wrong	Question 44: What is the sum of Goldman Sachs Group's Total Assets from 2011 to 2014 multiplied by the average of Operating Profit Margin between Air Products and Chemicals in 2015 and printing papers business of International Paper Company in 2005?
Wrong	Question 45: What is the Goldman Sachs Group's Total Assets change from 2014 to 2016 divided by the difference of Operating Profit Margin between Air Products and Chemicals in 2016 and printing papers business of International Paper Company in 2006?
Wrong	Question 46: What is the Goldman Sachs Group's Total Assets change from 2013 to 2017 divided by the sum of Operating Profit Margin between Air Products and Chemicals in 2016 and printing papers business of International Paper Company in 2004?
Wrong	Question 47: What is the difference of Union Pacific Corp's revenue from 2013 to 2016 divided by the sum of Operating Profit Margin between Air Products and Chemicals in 2015 and printing papers business of International Paper Company in 2004?
Wrong	Question 48: What is the average of Union Pacific Corp's revenue from 2013 to 2016 divided by the sum of Operating Profit Margin between Air Products and Chemicals in 2015 and printing papers business of International Paper Company in 2004?
Wrong	Question 49: What is the difference of Union Pacific Corp's revenue from 2013 to 2016 multiplied by the average of current ratio between American Tower Corp in 2012 and DISH Network Corporation in 2011?
Wrong	Question 13: What was the value in thousands of unvested restricted stock and performance awards at the weighted-average grant-date fair value as of December 31, 2018, of Global Payments?
Wrong	Question 50: What is the average of Union Pacific Corp's revenue from 2013 to 2016 divided by the average of Operating Profit Margin between Air Products and Chemicals in 2015 and printing papers business of International Paper Company in 2005?
Wrong	Question 32: What is the average of Operating Profit Margin between Air Products and Chemicals in 2016 and printing papers business of International Paper Company in 2005?
Wrong	Question 27: What is the International Paper Company's Operating Profit Margin of printing papers business in 2004?
Correct	Question 20: What is the percent of our network route miles that is owned rather than operated on pursuant to trackage rights or leases in 2016 Union Pacific Corp?
Correct	Question 3: What was the port call costsin of Royal Caribbean Cruises in 2012?
Correct	Question 4: What is the aggregate rent expense of American Tower Corp in 2014?
Correct	Question 5: What is the long-term component of BlackRock at 12/31/2011?
Correct	Question 6: At December 31, 2012, of The PNC Financial Services, what was the potential maximum exposure under the loss share arrangements?