Query Expansion

In [1]:
```
from helper_utils import load_chroma, word_wrap, project_embeddings
from chromadb.utils.embedding_functions import SentenceTransformerEmbeddingFunction
```

In [*]:
```
embedding_function = SentenceTransformerEmbeddingFunction()

chroma_collection = load_chroma(filename='microsoft_annual_report_2
```

In [3]:
```
import os
import openai
from openai import OpenAI

from dotenv import load_dotenv, find_dotenv
_ = load_dotenv(find_dotenv()) # read local .env file
openai.api_key = os.environ['OPENAI_API_KEY']

openai_client = OpenAI()
```

In [4]:
```
import umap

embeddings = chroma_collection.get(include=['embeddings'])['embedd
umap_transform = umap.UMAP(random_state=0, transform_seed=0).fit(em
projected_dataset_embeddings = project_embeddings(embeddings, umap_
```

```
/usr/local/lib/python3.9/site-packages/umap/umap_.py:1943: UserWarni
ng: n_jobs value -1 overridden to 1 by setting random_state. Use no
seed for parallelism.
  warn(f"n_jobs value {self.n_jobs} overridden to 1 by setting rando
m_state. Use no seed for parallelism.")
100%|██████████| 349/349 [08:06<00:00,  1.39s/it]
```

Expansion with generated answers

https://arxiv.org/abs/2305.03653 (https://arxiv.org/abs/2305.03653)

In [5]:
```
def augment_query_generated(query, model="gpt-3.5-turbo"):
    messages = [
        {
            "role": "system",
            "content": "You are a helpful expert financial researcl
        },
        {"role": "user", "content": query}
    ]

    response = openai_client.chat.completions.create(
        model=model,
        messages=messages,
    )
    content = response.choices[0].message.content
    return content
```

In [6]:
```
original_query = "Was there significant turnover in the executive
hypothetical_answer = augment_query_generated(original_query)

joint_query = f"{original_query} {hypothetical_answer}"
print(word_wrap(joint_query))
```

```
huggingface/tokenizers: The current process just got forked, after p
arallelism has already been used. Disabling parallelism to avoid dea
dlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLE
LISM=(true | false)

Was there significant turnover in the executive team? Yes, there was
significant turnover in the executive team during the fiscal year. F
our
key executives left the company, including the Chief Financial Offic
er,
Chief Operating Officer, Chief Marketing Officer, and Chief Technolo
gy
Officer. These departures were primarily driven by individual career
advancement opportunities and personal reasons. In order to maintain
continuity and ensure a smooth transition, the company promptly
appointed new executives with extensive experience in their respecti
ve
fields. The new team was carefully selected based on their expertise
and demonstrated leadership abilities, allowing the company to posit
ion
itself for continued growth and success in the market.
```

In [7]:
```
results = chroma_collection.query(query_texts=joint_query, n_resul
retrieved_documents = results['documents'][0]

for doc in retrieved_documents:
    print(word_wrap(doc))
    print('')
```

```
89 directors and executive officers of microsoft corporation directo
rs
satya nadella chairman and chief executive officer, microsoft
corporation sandra e. peterson 2, 3 operating partner, clayton,
dubilier & rice, llc john w. stanton 1, 4 founder and chairman, tril
ogy
partnerships reid g. hoffman 4 general partner, greylock partners pe
nny
s. pritzker 4 founder and chairman, psp partners, llc john w. thomps
on
3, 4 lead independent director, microsoft corporation hugh f. johnst
on
1 vice chairman and executive vice president and chief financial
officer, pepsico, inc. carlos a. rodriguez 1 chief executive office
r,
adp, inc. emma n. walmsley 2, 4 chief executive officer, gsk, plc te
ri
l. list 1, 3 former executive vice president and chief financial
officer, gap, inc. charles w. scharf 2, 3 chief executive officer an
d
president, wells fargo & company padmasree warrior 2 founder, presid
ent
and chief executive

the company engaged deloitte & touche llp, an independent registered
public accounting firm, to audit and render an opinion on the
consolidated financial statements and internal control over financia
l
reporting in accordance with the standards of the public company
accounting oversight board ( united states ). the board of director
s,
through its audit committee, consisting solely of independent direct
ors
of the company, meets periodically with management, internal auditor
s,
and our independent registered public accounting firm to ensure that
each is meeting its responsibilities and to discuss matters concerni
ng
internal controls and financial reporting. deloitte & touche llp and
the internal auditors each have full and free access to the audit
committee. satya nadella chief executive officer amy e. hood executi
ve
vice president and chief financial officer alice l. jolla corporate
vice president and chief accounting officer

officer, fable group inc. board committees 1. audit committee 2.
compensation committee 3. governance and nominating committee 4.
environmental, social, and public policy committee executive officer
s
satya nadella chairman and chief executive officer amy e. hood
executive vice president and chief financial officer judson althoff
executive vice president and chief commercial officer bradford l. sm
```

ith
vice chair and president christopher c. capossela executive vice
president, marketing and consumer business, and chief marketing offi
cer
christopher d. young executive vice president, business development,
strategy, and ventures kathleen t. hogan executive vice president an
d
chief human resources officer

88 report of independent registered public accounting firm to the
stockholders and the board of directors of microsoft corporation
opinion on internal control over financial reporting we have audited
the internal control over financial reporting of microsoft corporati
on
and subsidiaries (the “ company ”) as of june 30, 2022, based on
criteria established in internal control — integrated framework (20
13
) issued by the committee of sponsoring organizations of the treadwa
y
commission (coso). in our opinion, the company maintained, in all
material respects, effective internal control over financial reporti
ng
as of june 30, 2022, based on criteria establis hed in internal cont
rol
— integrated framework (2013) issued by coso. we have also audite
d,
in accordance with the standards of the public company accounting
oversight board (united states) (pcaob), the consolidated financ
ial
statements as of and for the year ended june 30, 2022, of the compan
y
and

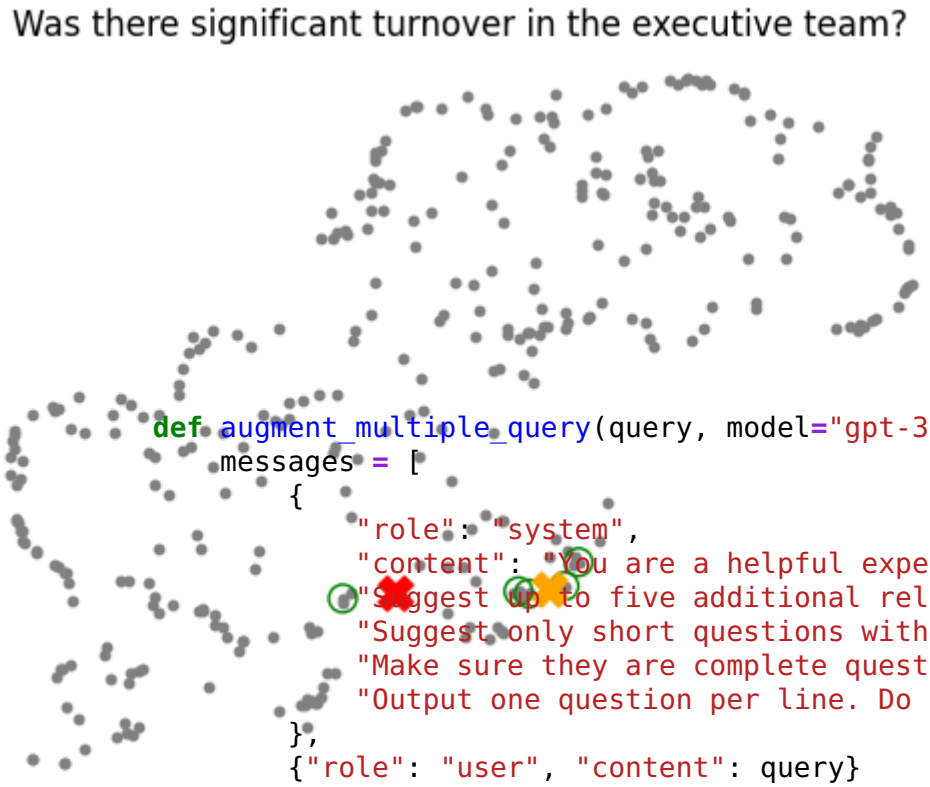
unresolved with the irs, evaluating management ’ s estimates relatin
g
to their determination of uncertain tax positions required extensive
audit effort and a high degree of auditor judgment, including
involvement of our tax specialists. how the critical audit matter wa
s
addressed in the audit our principal audit procedures to evaluate
management ’ s estimates of uncertain tax positions related to
unresolved transfer pricing issues included the following : • we
evaluated the appropriateness and consistency of management ’ s meth
ods
and assumptions used in the identification, recognition, measuremen
t,
and disclosure of uncertain tax positions, which included testing th
e
effectiveness of the related internal controls. • we read and evalua
ted
management ’ s documentation, including relevant accounting policies
and information obtained by management from outside tax specialists,
that detailed the basis of the uncertain tax positions.

In [8]:
```
retrieved_embeddings = results['embeddings'][0]
original_query_embedding = embedding_function([original_query])
augmented_query_embedding = embedding_function([joint_query])

projected_original_query_embedding = project_embeddings(original_qu
projected_augmented_query_embedding = project_embeddings(augmented_
projected_retrieved_embeddings = project_embeddings(retrieved_embed
```

```
100%|██████████| 1/1 [00:01<00:00,  1.24s/it]
100%|██████████| 1/1 [00:01<00:00,  1.22s/it]
100%|██████████| 5/5 [00:06<00:00,  1.31s/it]
```

In [9]:
```
import matplotlib.pyplot as plt

# Plot the projected query and retrieved documents in the embeddin
plt.figure()

#dataset embeddings plot ---> gray dots
plt.scatter(projected_dataset_embeddings[:, 0], projected_dataset_

#doc parts that were retrieved according to questiona and hypothet
plt.scatter(projected_retrieved_embeddings[:, 0], projected_retriev

#original query plotting--->red cross
plt.scatter(projected_original_query_embedding[:, 0], projected_ori

#query and hypothetical answer combined ---> yellow cross
plt.scatter(projected_augmented_query_embedding[:, 0], projected_au

plt.gca().set_aspect('equal', 'datalim')
plt.title(f'{original_query}')
plt.axis('off')
```

```
huggingface/tokenizers: The current process just got forked, after p
arallelism has already been used. Disabling parallelism to avoid dea
dlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLE
LISM=(true | false)
huggingface/tokenizers: The current process just got forked, after p
arallelism has already been used. Disabling parallelism to avoid dea
dlocks...
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLE
LISM=(true | false)

(-1.289832666516304, 8.499054208397865, 1.750054621696472, 9.1738760
23292541)
```



LM to generate the
nt document and then

```
def augment_multiple_query(query, model="gpt-3.5-turbo"):
    messages = [
        {
            "role": "system",
            "content": "You are a helpful expert financial researcl
            "Suggest up to five additional related questions to he
            "Suggest only short questions without compound sentence
            "Make sure they are complete questions, and that they a
            "Output one question per line. Do not number the questi
        },
        {"role": "user", "content": query}
    ]

    response = openai_client.chat.completions.create(
        model=model,
        messages=messages,
    )
    content = response.choices[0].message.content
    content = content.split("\n")
    return content
```

So the queries generated by the LLM will be:

In [11]:
```
original_query = "What were the most important factors that contril
augmented_queries = augment_multiple_query(original_query)

for query in augmented_queries:
    print(query)
```

```
What were the main drivers of revenue growth?
What were the key sources of revenue growth?
Were there any significant changes in revenue composition?
Did any particular product or service contribute significantly to th
e increase in revenue?
Were there any external factors that influenced the increase in reve
nue?
```

In [12]:
```
queries = [original_query] + augmented_queries

#passing all the queries to
results = chroma_collection.query(query_texts=queries, n_results=5

retrieved_documents = results['documents']

# Deduplicate the retrieved documents
unique_documents = set()
for documents in retrieved_documents:
    for document in documents:
        unique_documents.add(document)

for i, documents in enumerate(retrieved_documents):
    print(f"Query: {queries[i]}")
    print('')
    print("Results:")
    for doc in documents:
        print(word_wrap(doc))
        print('')
    print('-'*100)
```

```
Query: What were the most important factors that contributed to incr
eases in revenue?

Results:
engineering, gaming, and linkedin. • sales and marketing expenses
increased $ 1. 7 billion or 8 % driven by investments in commercial
sales and linkedin. sales and marketing included a favorable foreign
currency impact of 2 %. • general and administrative expenses increa
sed
$ 793 million or 16 % driven by investments in corporate functions.
operating income increased $ 13. 5 billion or 19 % driven by growth
across each of our segments. current year net income and diluted eps
were positively impacted by the net tax benefit related to the trans
fer
of intangible properties, which resulted in an increase to net incom
e
and diluted eps of $ 3. 3 billion and $ 0. 44, respectively. prior y
ear
net income and diluted eps were positively impacted by the net tax
```

In [13]:
```
original_query_embedding = embedding_function([original_query])
augmented_query_embeddings = embedding_function(augmented_queries)

project_original_query = project_embeddings(original_query_embeddi
project_augmented_queries = project_embeddings(augmented_query_embe
```

```
100%|██████████| 1/1 [00:01<00:00,  1.30s/it]
100%|██████████| 5/5 [00:06<00:00,  1.31s/it]
```

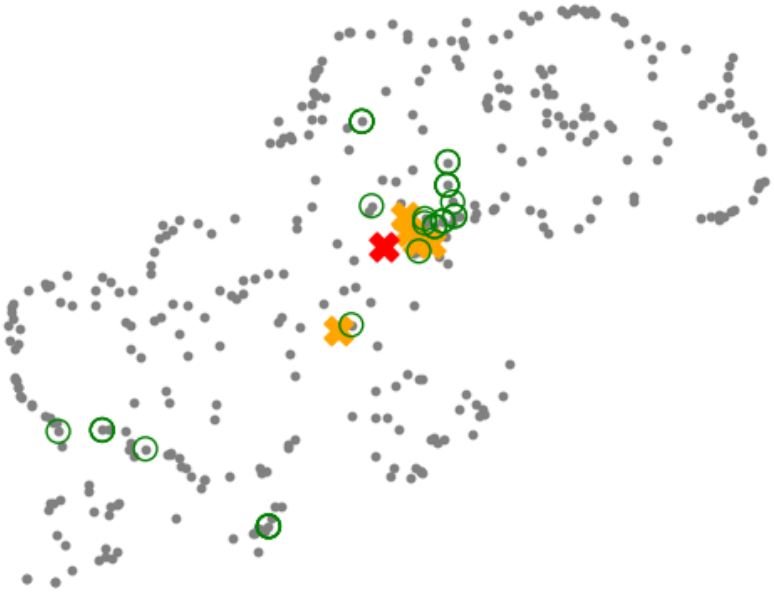
In [14]:
```
result_embeddings = results['embeddings']
result_embeddings = [item for sublist in result_embeddings for iter
projected_result_embeddings = project_embeddings(result_embeddings
```

```
100%|██████████| 30/30 [00:39<00:00,  1.33s/it]
```

In [15]:
```
import matplotlib.pyplot as plt

plt.figure()
plt.scatter(projected_dataset_embeddings[:, 0], projected_dataset_e
plt.scatter(project_augmented_queries[:, 0], project_augmented_que
plt.scatter(projected_result_embeddings[:, 0], projected_result_eml
plt.scatter(project_original_query[:, 0], project_original_query[:

plt.gca().set_aspect('equal', 'datalim')
plt.title(f'{original_query}')
plt.axis('off')
```

```
(-1.289832666516304, 8.499054208397865, 1.750054621696472, 9.1738760
23292541)
```

What were the most important factors that contributed to increases in revenue?



In [ ]:

In [ ]: