

✓ Bridging the Gaps in LLM Interpretability of Unstructured Financial Earnings Data with Knowledge Graphs

- Use case: answer questions about companies' financial performance based on the transcripts of their earnings calls.
- Uses StrictJSON to parse the Knowledge Graph: <https://github.com/tanchongmin/strictjson>

> Import packages

🔍 5 cells hidden

✓ Load dataset

```
# Function to parse JSONL file
def parse_jsonl(file_path):
    data = []
    with open(file_path, 'r') as file:
        for line in file:
            data.append(json.loads(line))
    return data

file_path = 'data/earnings-transcripts.jsonl'
earnings_data = parse_jsonl(file_path)
```

```
df = pd.DataFrame(earnings_data)
```

```
df.head()
```

	date	exchange	q	ticker	transcript
0	Aug 27, 2020, 9:00 p.m. ET	NASDAQ: BILI	2020-Q2	BILI	Prepared Remarks:\nOperator\nGood day, and wel...
1	Jul 30, 2020, 4:30 p.m. ET	NYSE: GFF	2020-Q3	GFF	Prepared Remarks:\nOperator\nThank you for sta...
2	Oct 23, 2019, 5:00 p.m. ET	NASDAQ: LRCX	2020-Q1	LRCX	Prepared Remarks:\nOperator\nGood day and welc...
3	Nov 6, 2019, 12:00 p.m. ET	NASDAQ: BBSI	2019-Q3	BBSI	Prepared Remarks:\nOperator\nGood day, everyon...
4	Aug 7, 2019, 8:30 a.m. ET	NASDAQ: CSTE	2019-Q2	CSTE	Prepared Remarks:\nOperator\nGreetings and wel...

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype
---  ---
 0   date        1000 non-null   object
 1   exchange    1000 non-null   object
 2   q           1000 non-null   object
 3   ticker      1000 non-null   object
 4   transcript  1000 non-null   object
dtypes: object(5)
memory usage: 39.2+ KB
```

```
df.describe()
```



	date	exchange	q	ticker	transcript
count	1000	1000	1000	1000	1000
unique	783	703	22	703	962
top	May 27, 2020, 9:00 p.m. ET	NASDAQ: TSLA	2020- Q4	TSLA	Prepared Remarks:\nOperator\nGood morning, and...
freq	7	8	210	8	7

```
# Remove data and exchange columns as they are not needed
df = df[['q', 'ticker', 'transcript']]
df.head()
```




	q	ticker	transcript
0	2020-Q2	BILI	Prepared Remarks:\nOperator\nGood day, and wel...
1	2020-Q3	GFF	Prepared Remarks:\nOperator\nThank you for sta...
2	2020-Q1	LRCX	Prepared Remarks:\nOperator\nGood day and welc...
3	2019-Q3	BBSI	Prepared Remarks:\nOperator\nGood day, everyon...
4	2019-Q2	CSTE	Prepared Remarks:\nOperator\nGreetings and wel...

```
# Create new df only with rows with specific tickers
tickers = ['TSM', 'COHR', 'SWKS', 'ASML', 'MTSI']
df_new = df[df['ticker'].isin(tickers)]
df_new
```



	q	ticker	transcript
15	2020-Q4	TSM	Prepared Remarks:\nJeff Su -- Director of Inve...
136	2023-Q1	COHR	Prepared Remarks:\nOperator\nLadies and gentle...
153	2022-Q3	TSM	Prepared Remarks:\nJeff Su\nGood afternoon, ev...
225	2021-Q4	MTSI	Prepared Remarks:\nOperator\nWelcome to MACOM'...
506	2021-Q4	MTSI	Prepared Remarks:\nOperator\nWelcome to MACOM'...
511	2023-Q2	COHR	Prepared Remarks:\nOperator\nGood day, and tha...
591	2019-Q4	MTSI	Prepared Remarks:\nOperator\nGood afternoon, a...
596	2023-Q2	COHR	Prepared Remarks:\nOperator\nGood day, and tha...
620	2020-Q2	SWKS	Prepared Remarks:\nOperator\nGood afternoon an...
924	2022-Q1	SWKS	Prepared Remarks:\nOperator\nGood afternoon, a...
936	2020-Q1	ASML	Prepared Remarks:\nOperator\nThank you for sta...

```
# Count words for each transcript
df_new['word_count'] = df_new['transcript'].apply(lambda x: len(x.split()))
df_new
```

 /var/folders/xb/h3ltfn611fg9ycm0s6vn5qpw0000gn/T/ipykernel_15472/1508092007.py:2
 A value is trying to be set on a copy of a slice from a DataFrame.
 Try using `.loc[row_indexer,col_indexer] = value` instead

See the caveats in the documentation: <https://pandas.pydata.org/pandas-docs/stab>
`df_new['word_count'] = df_new['transcript'].apply(lambda x: len(x.split()))`

	q	ticker	transcript	word_count
15	2020-Q4	TSM	Prepared Remarks:\nJeff Su -- Director of Inve...	12478
136	2023-Q1	COHR	Prepared Remarks:\nOperator\nLadies and gentle...	11602
153	2022-Q3	TSM	Prepared Remarks:\nJeff Su\nGood afternoon, ev...	10069
225	2021-Q4	MTSI	Prepared Remarks:\nOperator\nWelcome to MACOM'...	11095
506	2021-Q4	MTSI	Prepared Remarks:\nOperator\nWelcome to MACOM'...	11095
511	2023-Q2	COHR	Prepared Remarks:\nOperator\nGood day, and tha...	10960
591	2019-Q4	MTSI	Prepared Remarks:\nOperator\nGood afternoon, a...	7136
596	2023-Q2	COHR	Prepared Remarks:\nOperator\nGood day, and tha...	10960
620	2020-Q2	SWKS	Prepared Remarks:\nOperator\nGood afternoon an...	8089
924	2022-Q1	SWKS	Prepared Remarks:\nOperator\nGood afternoon, a...	7562
936	2020-Q1	ASML	Prepared Remarks:\nOperator\nThank you for sta...	9332

> Utility functions

[] ↪ 1 cell hidden

> Data Preprocessing

[] ↪ 8 cells hidden

✓ Build Knowledge Graph

```
def build_knowledge_graph(transcript_summary):
    '''Extract knowledge graph from summarized text using schema'''
    completion = openai.chat.completions.create(
        model="gpt-4-turbo",
        response_format={ "type": "json_object" },
        messages=[
            {"role": "system", "content": '''You are a knowledge graph builder, extract nodes and edges for a knowle
            You are to output relations between two objects in the form (object_1, relation, object_2).
            All information about dates must be included.
            Example Input: John bought a laptop
            Example Output: [('John', 'bought', 'laptop')]
            Example Input: John built a house in 2019
            Example Output: [('John', 'built', 'house'), ('house', 'built in', '2019')]
            The final output should be in JSON as follows: {"List of triplets": "List of triplets of the form (objec
            {"role": "user", "content": f"Here's the text: {transcript_summary}"}
        ]
    )

    answer = json.loads(completion.choices[0].message.content) if completion.choices else "No response"

    return answer
```

```
def plot_graph(kg):
    ''' Plots graph based on knowledge graph '''
    # Create graph
    G = nx.DiGraph()
    G.add_edges_from((source, target, {'relation': relation}) for source, relation, target in kg)

    # Plot the graph
    plt.figure(figsize=(10,6), dpi=300)
    pos = nx.spring_layout(G, k=3, seed=0)

    nx.draw_networkx_nodes(G, pos, node_size=1500)
    nx.draw_networkx_edges(G, pos, edge_color='gray')
    nx.draw_networkx_labels(G, pos, font_size=12)
    edge_labels = nx.get_edge_attributes(G, 'relation')
    nx.draw_networkx_edge_labels(G, pos, edge_labels=edge_labels, font_size=10)

    # Display the plot
    plt.axis('off')
    plt.show()

# Build knowledge graph for each transcript clean summary
df_earnings['knowledge_graph'] = df_earnings['clean_summary'].apply(build_knowledge_graph)

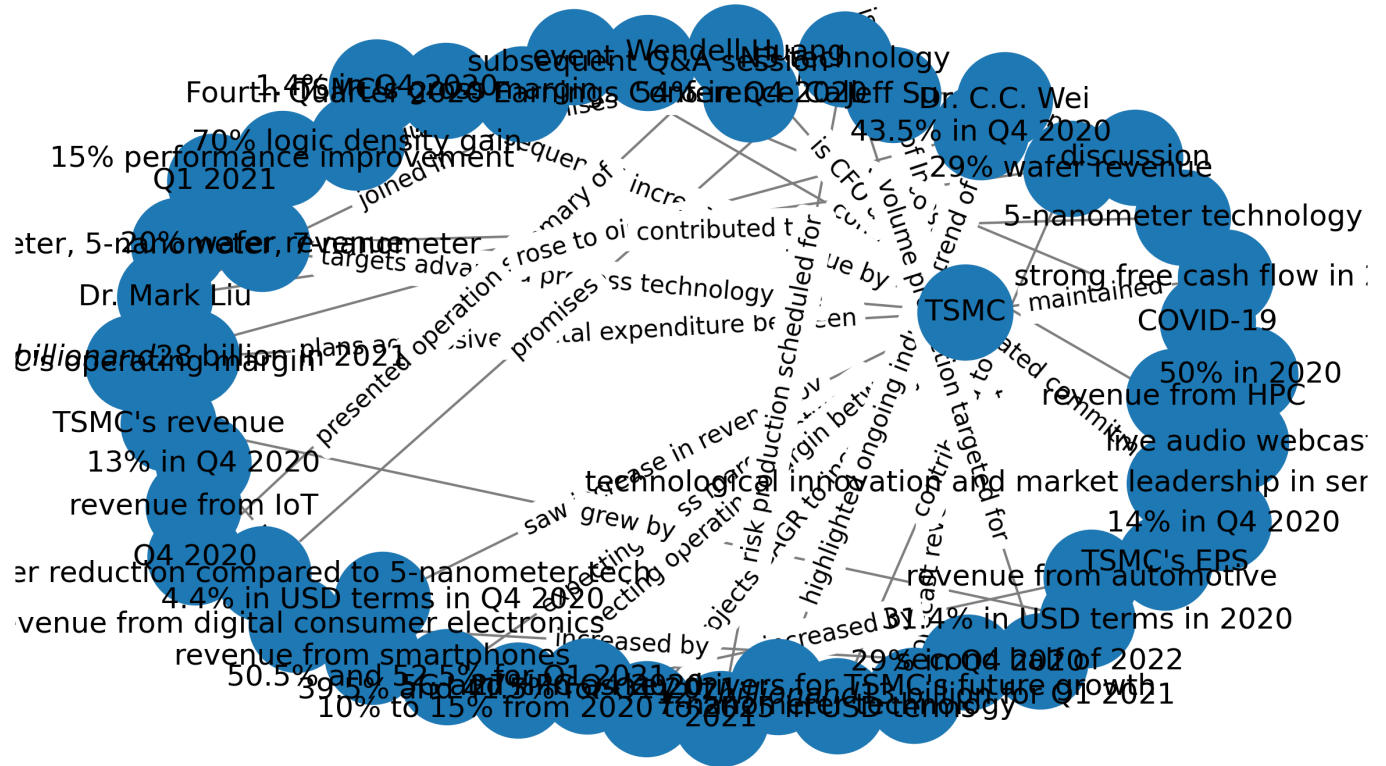
# Save df_earnings to csv to avoid re-building knowledge graph
df_earnings.to_csv('data/earnings-transcripts-kg.csv', index=False)
```

```
df_earnings.head()
```



	q	ticker	transcript	word_count	summary	summary_word
0	2020-Q4	TSM	Prepared Remarks:\nJeff Su -- Director of Inve...	12478	Prepared Remarks:\nJeff Su, Director of Invest...	
1	2023-Q1	COHR	Prepared Remarks:\nOperator\nLadies and gentle...	11602	Operator\nLadies and gentlemen, welcome to the...	
2	2022-Q3	TSM	Prepared Remarks:\nJeff Su\nGood afternoon, ev...	10069	Good afternoon, everyone, and welcome to TSMC'...	
3	2021-Q4	MTSI	Prepared Remarks:\nOperator\nWelcome to MACOM'...	11095	MACOM Technology Solutions Holdings, Inc. (MAC...	
4	2021-Q4	MTSI	Prepared Remarks:\nOperator\nWelcome to MACOM'...	11095	Welcome to the summary of MACOM's Fourth	

```
# Plot knowledge graph for the first transcript
kg = df_earnings['knowledge_graph'][0]['List of triplets']
plot_graph(kg)
```



▼ Parse Knowledge Graph

```
# def parse_knowledge_graph(kg, question):
#     '''Parse knowledge graph to extract relevant relations'''
#     completion = openai.chat.completions.create(
#         model="gpt-4-turbo",
#         response_format={ "type": "json_object" },
#         messages=[
#             {
#                 "role": "system", "content": f'''You are a knowledge graph parser for the following knowledge graph {kg}
#                 Only output the triplets that are relevant to the question.
#                 The final output should be in JSON as follows: {"Parsed Knowledge Graph": "List of triplets of the form
#                 {"role": "user", "content": f"Here's the question: {question}"}}
#             ]
#         )
#     )

#     answer = json.loads(completion.choices[0].message.content) if completion.choices else "No response"

#     return answer
```

```
triplets_text = [{"{} {} {}".format(src, rel, trg) for src, rel, trg in kg}]
```

```
vectorizer = TfidfVectorizer()
tfidf_matrix = vectorizer.fit_transform(triplets_text)
```

```
question = "What is the expected revenue growth rate for TSMC from 2020 to 2025 in US dollar terms"
query_vector = vectorizer.transform([question])
similarity_scores = cosine_similarity(query_vector, tfidf_matrix)[0]
print(similarity_scores)
```

```
[0.15945388 0.034025 0. 0.07915075 0.20671698 0.04059611
 0.07175265 0.04445658 0.03914005 0.03847574 0.03487279 0.21276046
 0.33820155 0.19127015 0.19532939 0.13950886 0.14278155 0.23790757
 0.22952955 0.23270557 0.23270557 0.19645593 0.32770995 0.13745994]
```

```
0.10694529 0.12973258 0.10069001 0.09927561 0.04732378 0.02660869
0.18420565 0.48020193 0. 0. 0.0589088 0.08288286
0.06398596 0.10530552]
```

```
import numpy as np
```

```
threshold = 0.2
```

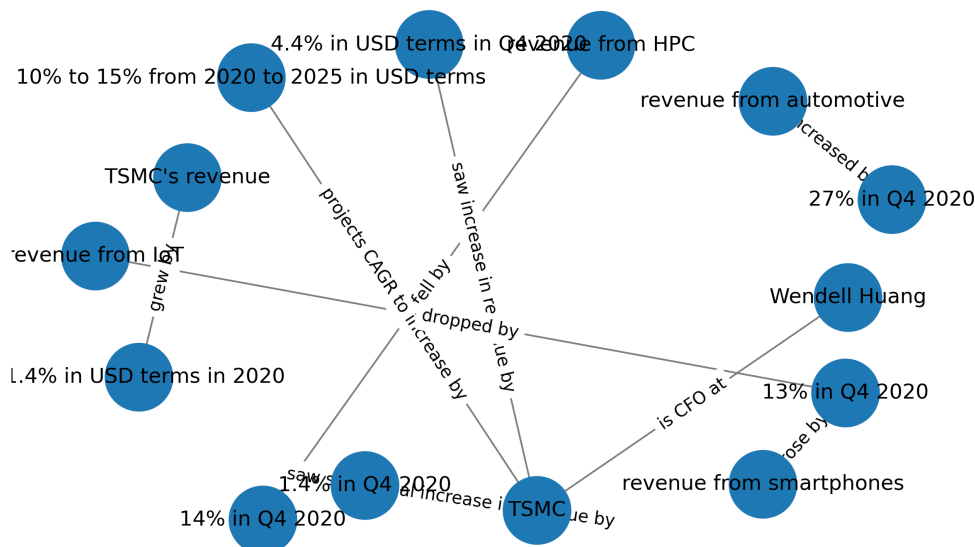
```
relevant_indices = np.where(similarity_scores > threshold)[0]
```

```
relevant_triplets = [kg[i] for i in relevant_indices]
```

```
print(f'Relevant Triplets: {relevant_triplets}')
```

```
➦ Relevant Triplets: [['Wendell Huang', 'is CFO at', 'TSMC'], ['TSMC', 'saw sequential increase in revenue by', '1
```

```
# Plot knowledge graph based on relevant triplets
plot_graph(relevant_triplets)
```



▼ Generate Answer

```
# Ground truth answer
```

```
answer_from_QA = "The expected revenue growth rate for TSMC from 2020 to 2025 in US dollar terms is 10% to 15% CAGR."
```

```
print('Question:', question)
```

```
print('Answer from Parsed Knowledge Graph:', chat(f'''Use the knowledge graph to answer the following question.
```

```
If you are unsure, output 'No Info'
```

```
Knowledge Graph: {relevant_triplets}''',
```

```
user_prompt = f'''Question: {question}''')
```

```
print('Ground Truth Answer:', answer_from_QA)
```



```
Question: What is the expected revenue growth rate for TSMC from 2020 to 2025 in US dollar terms
```

```
Answer from Parsed Knowledge Graph: TSMC projects CAGR to increase by 10% to 15% from 2020 to 2025 in USD terms.
```

Ground Truth Answer: The expected revenue growth rate for TSMC from 2020 to 2025 in US dollar terms is 10% to 15%

✓ Eval

```
!pip install datasets
```

```

Collecting datasets
  Downloading datasets-2.19.1-py3-none-any.whl (542 kB)
    542.0/542.0 kB 10.6 MB/s eta 0:00:00
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from datasets) (3.14.0)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (from datasets) (1.25.2)
Requirement already satisfied: pyarrow>=12.0.0 in /usr/local/lib/python3.10/dist-packages (from datasets) (14.0.1)
Requirement already satisfied: pyarrow-hotfix in /usr/local/lib/python3.10/dist-packages (from datasets) (0.6)
Requirement already satisfied: dill<0.3.9,>=0.3.0 in /usr/local/lib/python3.10/dist-packages (from datasets) (0.3.7)
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages (from datasets) (2.0.3)
Requirement already satisfied: requests>=2.19.0 in /usr/local/lib/python3.10/dist-packages (from datasets) (2.31.0)
Requirement already satisfied: tqdm>=4.62.1 in /usr/local/lib/python3.10/dist-packages (from datasets) (4.66.4)
Collecting xxhash (from datasets)
  Downloading xxhash-3.4.1-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
    194.1/194.1 kB 27.1 MB/s eta 0:00:00
Collecting multiprocessing (from datasets)
  Downloading multiprocessing-0.70.16-py310-none-any.whl (134 kB)
    134.8/134.8 kB 20.2 MB/s eta 0:00:00
Requirement already satisfied: fsspec[http]<=2024.3.1,>=2023.1.0 in /usr/local/lib/python3.10/dist-packages (from datasets) (2024.3.1)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.10/dist-packages (from datasets) (3.9.5)
Collecting huggingface-hub>=0.21.2 (from datasets)
  Downloading huggingface_hub-0.23.0-py3-none-any.whl (401 kB)
    401.2/401.2 kB 43.5 MB/s eta 0:00:00
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from datasets) (24.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from datasets) (6.0.1)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (23.2.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (6.0.5)
Requirement already satisfied: yarl<2.0,>=1.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (1.9.7)
Requirement already satisfied: async-timeout<5.0,>=4.0 in /usr/local/lib/python3.10/dist-packages (from aiohttp->datasets) (4.0.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub) (4.9.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->datasets) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->datasets) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->datasets) (2.2.3)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->datasets) (2024.2.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.10/dist-packages (from pandas->datasets) (2.9.0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas->datasets) (2024.1)
Requirement already satisfied: tzdata>=2022.1 in /usr/local/lib/python3.10/dist-packages (from pandas->datasets) (2024.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil->datasets) (1.16.0)
Installing collected packages: xxhash, multiprocessing, huggingface-hub, datasets
  Attempting uninstall: huggingface-hub
    Found existing installation: huggingface-hub 0.20.3
    Uninstalling huggingface-hub-0.20.3:
      Successfully uninstalled huggingface-hub-0.20.3
Successfully installed datasets-2.19.1 huggingface-hub-0.23.0 multiprocessing-0.70.16 xxhash-3.4.1

```

```

import pandas as pd
df_earnings = pd.read_csv('earnings-transcripts-kg.csv')
df_earnings.sort_values(by='ticker', inplace=True)
df_earnings.head()

```



	q	ticker	transcript	word_count	summary	summary_w
10	2020-Q1	ASML	Remarks:\nOperator\nThank you for sta... Prepared	9332	ASML's 2020 first-quarter earnings call, held ...	
1	2023-Q1	COHR	Remarks:\nOperator\nLadies and gentle... Prepared	11602	Operator\nLadies and gentlemen, welcome to the...	
5	2023-Q2	COHR	Remarks:\nOperator\nGood day, and tha... Prepared	10960	Operator\nWelcome to Coherent Corp.'s FY '23 ...	
7	2023-Q2	COHR	Remarks:\nOperator\nGood day, and tha... Prepared	10960	Operator\nWelcome to the Coherent Corp. FY '23...	
3	2021-Q4	MTSI	Remarks:\nOperator\nWelcome to MACOM'... Prepared	11095	MACOM Technology Solutions Holdings, Inc. (MAC...	

```
from datasets import load_dataset
dataset = load_dataset("lamini/earnings-calls-qa")
print(dataset['train'][0])
```



```
/usr/local/lib/python3.10/dist-packages/huggingface_hub/utils/_token.py:89: UserWarning: The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (.
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public datasets.
warnings.warn(
Downloading readme: 100% 1.06k/1.06k [00:00<00:00, 94.5kB/s]
Downloading data: 100% 3.89G/3.89G [03:06<00:00, 20.5MB/s]
Generating train split: 100% 860164/860164 [00:10<00:00, 80049.77 examples/s]
{'question': "What was TSMC's revenue in US dollar terms in 2019 ", 'answer': '

```

```
df_groundtruth = pd.DataFrame(dataset['train'])
```

```
df_groundtruth
```




	question	answer	date	transcript	q	ticker	predictions
0	What was TSMC's revenue in US dollar terms in ...	I do not know. The transcript does not provid...	Jan 13, 2022, 1:00 a.m. ET	and for our industry-leading advanced and spec...	2021-Q4	TSM	[{'class_id': 0, 'class_name': 'correct', 'pro...
1	What was TSMC's EPS in 2019	I do not know. The transcript does not provid...	Jan 13, 2022, 1:00 a.m. ET	and for our industry-leading advanced and spec...	2021-Q4	TSM	[{'class_id': 0, 'class_name': 'correct', 'pro...
2	What was TSMC's capex spending in 2019	I do not know. The transcript does not provid...	Jan 13, 2022, 1:00 a.m. ET	and for our industry-leading advanced and spec...	2021-Q4	TSM	[{'class_id': 0, 'class_name': 'correct', 'pro...
3	What is the expected growth rate of global sma...	The expected growth rate of global smartphone...	Jan 14, 2021, 1:00 a.m. ET	g demand for our advanced technologies in the ...	2020-Q4	TSM	[{'class_id': 0, 'class_name': 'correct', 'pro...
4	What is the expected penetration rate for 5G s...	The expected penetration rate for 5G smartpho...	Jan 14, 2021, 1:00 a.m. ET	g demand for our advanced technologies in the ...	2020-Q4	TSM	[{'class_id': 0, 'class_name': 'correct', 'pro...
...
860159	What was the company's adjusted EBITDA in 2021...	The company's adjusted EBITDA in 2021-Q1 was ...	May 06, 2021, 4:30 p.m. ET	but including restricted cash, total cash was ...	2021-Q1	OTRK	[{'class_id': 0, 'class_name': 'correct', 'pro...

```
df_earnings[df_earnings['ticker'] == 'TSM']['transcript'][0]
```



'Prepared Remarks:\nJeff Su -- Director of Investor Relations\n[Foreign Speech]
Good afternoon, everyone, and welcome to TSMC's Fourth Quarter 2020 Earnings Conference Call. This is Jeff Su, TSMC's Director of Investor Relations and your host for today.\nTo prevent the spread of COVID-19, TSMC is hosting our earnings conference call via live audio webcast through the Company's website at www.tsmc.com, where you can also download the earnings release materials. If you are joining us through the conference call, your dial-in lines are in listen-only mode.\nThe format for today's event will be as follows. First, TSMC's Vice President and CFO, Mr. Wendell Huang, will summarize our operations in the fourth quarter 2020, followed by our guidance for the first quarter 2021. Afterwards, Mr. Hua

```
filtered_df = df_groundtruth[(df_groundtruth['ticker'] == 'TSM') & (df_groundtruth['transcript'].str.contains('spread'))]
filtered_df
```



	question	answer	date	transcript	q	ticker	predictions
167299	What was the revenue increase in NT dollars fo...	The revenue increase in NT dollars for the se...	Jul 14, 2022, 2:00 a.m. ET	Prepared Remarks:\nJeff Su\n[Foreign language]...	2022-Q2	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
167300	What was the gross margin increase in percenta...	The gross margin increase in percentage point...	Jul 14, 2022, 2:00 a.m. ET	Prepared Remarks:\nJeff Su\n[Foreign language]...	2022-Q2	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
167301	What was the operating margin increase in perc...	The operating margin increase in percentage p...	Jul 14, 2022, 2:00 a.m. ET	Prepared Remarks:\nJeff Su\n[Foreign language]...	2022-Q2	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
167302	What was the revenue increase in NT dollars fo...	The revenue increase in NT dollars for the se...	Jul 14, 2022, 2:00 a.m. ET	Prepared Remarks:\nJeff Su\n[Foreign language]...	2022-Q2	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
167303	What was the gross margin increase in percenta...	The gross margin increase in percentage point...	Jul 14, 2022, 2:00 a.m. ET	Prepared Remarks:\nJeff Su\n[Foreign language]...	2022-Q2	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
167304	What was the operating margin increase in perc...	The operating margin increase in percentage p...	Jul 14, 2022, 2:00 a.m. ET	Prepared Remarks:\nJeff Su\n[Foreign language]...	2022-Q2	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
185791	What was the revenue increase in NT dollars an...	The revenue increase in NT dollars for the fi...	Apr 15, 2021, 2:00 a.m. ET	Prepared Remarks:\nJeff Su -- Director of Inve...	2021-Q1	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
185792	What was the gross margin in the first quarter...	The gross margin in the first quarter 2021 wa...	Apr 15, 2021, 2:00 a.m. ET	Prepared Remarks:\nJeff Su -- Director of Inve...	2021-Q1	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
185793	What was the revenue contribution by platform ...	The revenue contribution by platform in the f...	Apr 15, 2021, 2:00 a.m. ET	Prepared Remarks:\nJeff Su -- Director of Inve...	2021-Q1	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
241033	What was the revenue contribution of the autom...	The revenue contribution of the automotive pl...	Jan 13, 2022, 1:00 a.m. ET	Prepared Remarks:\nJeff Su\n[Foreign language]...	2021-Q4	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
241034	What was the increase in current liabilities i...	The increase in current liabilities in the fo...	Jan 13, 2022, 1:00 a.m. ET	Prepared Remarks:\nJeff Su\n[Foreign language]...	2021-Q4	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
241035	What was the revenue contribution of 5-nanomet...	The revenue contribution of 5-nanometer proce...	Jan 13, 2022, 1:00 a.m. ET	Prepared Remarks:\nJeff Su\n[Foreign language]...	2021-Q4	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
285937	What was the revenue contribution of smartphon...	The revenue contribution of smartphone in the...	Jan 14, 2021, 1:00 a.m. ET	Prepared Remarks:\nJeff Su -- Director of Inve...	2020-Q4	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
285938	What was the gross margin in the fourth quarte...	The gross margin in the fourth quarter of 202...	Jan 14, 2021, 1:00 a.m. ET	Prepared Remarks:\nJeff Su -- Director of Inve...	2020-Q4	TSM	{'class_id': 0, 'class_name': 'correct', 'pro...
285939	What was the revenue contribution of 5-	The revenue contribution of 5-	Jan 14, 2021 1:00	Prepared Remarks:\nJeff	2020-	TSM	{'class_id': 0, 'class_name': 'correct'

```
filtered_df = filtered_df.reset_index(drop=True)
filtered_df['transcript'][0]
```

'Prepared Remarks:\nJeff Su\n[Foreign language] Good afternoon, everyone, and welcome to TSMC's second quarter 2022 earnings conference call. This is Jeff Su, TSMC's director of investor relations and your host for today. To prevent the spread of COVID-19, TSMC is hosting our earnings conference call via live audio webcast through the company's website at www.tsmc.com, where you can also download the earnings release materials. [Operator instructions] The format for today's event will be as follows. First, TSMC's vice president and CFO, Mr.\nWendell Huang, will summarize our operations in the second quarter 2022, followed by our guidance for the third quarter 2022. Afterwards, Mr. Huang and TSMC's CEO, Dr. C.C.\nWei, will jointly provide the company's key messages. Then TSMC's chairman, Dr. Mark Liu, will host the Q&A session, where all three executives will entertain your questions. As usual, I would like to remind everybody that today's discussions may contain forward-looking statements.'

```
df_earnings_2 = df_earnings.copy()
df_earnings_2 = df_earnings_2.merge(df_groundtruth[['ticker', 'q', 'question', 'answer']],
                                   on=['ticker', 'q'], how='left')
```

```
df_earnings_2
```



	Q1	ASML	Remarks:\nOperator\nThank you for sta...	9332	earnings call, held ...
4	2020-Q1	ASML	Prepared Remarks:\nOperator\nThank you for sta...	9332	ASML's 2020 first-quarter earnings call, held ...
...
1585	2022-Q3	TSM	Prepared Remarks:\nJeff Su\nGood afternoon, ev...	10069	Good afternoon, everyone, and welcome to TSMC'...
1586	2022-Q3	TSM	Prepared Remarks:\nJeff Su\nGood afternoon, ev...	10069	Good afternoon, everyone, and welcome to TSMC'...
1587	2022-Q3	TSM	Prepared Remarks:\nJeff Su\nGood afternoon, ev...	10069	Good afternoon, everyone, and welcome to TSMC'...
1588	2022-Q3	TSM	Prepared Remarks:\nJeff Su\nGood afternoon, ev...	10069	Good afternoon, everyone, and welcome to TSMC'...
1589	2022-Q3	TSM	Prepared Remarks:\nJeff Su\nGood afternoon, ev...	10069	Good afternoon, everyone, and welcome to TSMC'...

1590 rows x 10 columns

```
df_earnings_2['ticker'].value_counts()
```

```
↗ ticker
COHR    708
MTSI    402
TSM     276
SWKS    132
ASML     72
Name: count, dtype: int64
```

```
import json
import ast
```

```
def match_string(string):
    match = re.search(r"'List of triplets': ([.*\s]*)", string)
    if match:
        # The found list as a string
        list_str = match.group(1)

        try:
            # Convert the string representation of the list to an actual list
            list_of_triplets = ast.literal_eval(list_str)
            # print(list_of_triplets)
            return list_of_triplets
        except SyntaxError as e:
            print("Error converting string to list:", e)
    else:
        return None
```

```
def process_ticker_data(df, tickers):
    results = []
    vectorizer = TfidfVectorizer()

    # Process each ticker
    for ticker in tickers:
        # Filter DataFrame for the ticker and limit to first 5 rows
        ticker_df = df[df['ticker'] == ticker]

        for index, row in ticker_df.iterrows():
            # Extract the knowledge graph for the current row
            kg = match_string(row['knowledge_graph'])
            # kg = knowledge_graph.get('List of triplets', [])
            # print(kg)
            questions = row['question']
            answers = row['answer']
            triplets_text = [{"src", "rel", "trg"} for src, rel, trg in kg]

            # Compute TF-IDF matrix for knowledge graph triplets
            tfidf_matrix = vectorizer.fit_transform(triplets_text)
            query_vector = vectorizer.transform([questions])
            similarity_scores = cosine_similarity(query_vector, tfidf_matrix)[0]

            # Determine relevant triplets based on a threshold
            threshold = 0.2
            relevant_indices = np.where(similarity_scores > threshold)[0]
            relevant_triplets = [kg[i] for i in relevant_indices]
            print(relevant_triplets)
            # Generate output
            parsed_answer = (chat(f'''Use the knowledge graph to answer the following question.
                                If you are unsure, output 'No Info'
                                Knowledge Graph: {relevant_triplets}''',
```

14/17

```
[['TSMC', 'saw sequential increase in revenue by', '1.4% in Q4 2020'], ['TSMC', 'saw increase in revenue by',
[['TSMC', 'saw sequential increase in revenue by', '1.4% in Q4 2020'], ['TSMC', 'saw increase in revenue by',
[['Wendell Huang', 'is CFO at', 'TSMC'], ['Wendell Huang', 'followed by guidance presentation for', 'Q1 2021']
[['Wendell Huang', 'is CFO at', 'TSMC'], ['TSMC', 'saw increase in revenue by', '4.4% in USD terms in Q4 2020']
[['Wendell Huang', 'is CFO at', 'TSMC'], ['TSMC', 'saw increase in revenue by', '4.4% in USD terms in Q4 2020']
```

for result in results:

```
print('Question:', result['Question'])
print('Parsed Answer:', result['Parsed Answer'])
print('Ground Truth Answer:', result['Ground Truth Answer'])
```

```

Question: What is the expected growth rate of global smartphone units in 2021
Parsed Answer: No Info
Ground Truth Answer: The expected growth rate of global smartphone units in 2021 is 10%.
Question: What is the expected penetration rate for 5G smartphones in the total smartphone market in 2021
Parsed Answer: No Info
Ground Truth Answer: The expected penetration rate for 5G smartphones in the total smartphone market in 2021
Question: What is the expected revenue growth rate for TSMC from 2020 to 2025 in US dollar terms
Parsed Answer: TSMC projects CAGR to increase by 10% to 15% from 2020 to 2025 in USD terms.
Ground Truth Answer: The expected revenue growth rate for TSMC from 2020 to 2025 in US dollar terms is 10% t
Question: What is the expected revenue from TSMC's back-end services, which include InFO's advanced packaging
Parsed Answer: No Info
Ground Truth Answer: The expected revenue from TSMC's back-end services, which include InFO's advanced packa
Question: What is the expected full volume production of SoIC in 2022
Parsed Answer: No Info
Ground Truth Answer: The expected full volume production of SoIC in 2022 is targeted.
Question: What is the expected adoption of HPC applications for SoIC in 2022
Parsed Answer: No Info
Ground Truth Answer: The expected adoption of HPC applications for SoIC in 2022 is not explicitly stated in
Question: What is the expected growth rate for the HPC platform in 2021
Parsed Answer: No Info
Ground Truth Answer: The expected growth rate for the HPC platform in 2021 is not explicitly stated in the t
Question: What is the expected growth rate for the automotive platform in 2021
Parsed Answer: No Info
Ground Truth Answer: The expected growth rate for the automotive platform in 2021 is not provided in the tra
Question: What is the expected growth rate for the smartphone platform in 2021
Parsed Answer: No Info
Ground Truth Answer: The expected growth rate for the smartphone platform in 2021 is similar to the corporat
Question: What was TSMC's revenue in US dollar terms in 2020
Parsed Answer: TSMC's revenue grew by 31.4% in USD terms in 2020.
Ground Truth Answer: I do not know. The transcript does not provide the revenue of TSMC in US dollar terms f
Question: What was the growth rate of TSMC's revenue in US dollar terms in 2020 compared to the previous year
Parsed Answer: TSMC's revenue grew by 31.4% in USD terms in 2020.
Ground Truth Answer: The growth rate of TSMC's revenue in US dollar terms in 2020 compared to the previous y
Question: What is TSMC's revenue growth forecast for the full-year of 2021 in US dollar terms
Parsed Answer: No Info
Ground Truth Answer: TSMC's revenue growth forecast for the full-year of 2021 in US dollar terms is not expl
Question: What is the number of transistors per millimeter square at 5-nanometer
Parsed Answer: No Info
Ground Truth Answer: The number of transistors per millimeter square at 5-nanometer is 175 million.
Question: What is the number of transistors per millimeter square at 3-nanometer
Parsed Answer: No Info
Ground Truth Answer: I do not know the number of transistors per millimeter square at 3-nanometer from the g
Question: What is the expected reduction in capex after improving EUV productivity to the optimized level
Parsed Answer: No Info
Ground Truth Answer: The expected reduction in capex after improving EUV productivity to the optimized level
Question: What is the company's target CAGR for the next five years
Parsed Answer: No Info
Ground Truth Answer: The company's target CAGR for the next five years is between 10% to 15%.
Question: What is the company's target CAGR for the next five years based on a 2020 high?
Parsed Answer: TSMC projects CAGR to increase by 10% to 15% from 2020 to 2025 in USD terms.
Ground Truth Answer: The company's target CAGR for the next five years based on a 2020 high is between 10% t
Question: What is the company's target CAGR for the next five years based on a 2020 high and a 10% to 15% ran
Parsed Answer: TSMC projects CAGR to increase by 10% to 15% from 2020 to 2025 in USD terms.
Ground Truth Answer: The company's target CAGR for the next five years based on a 2020 high and a 10% to 15%
Question: What is the capex guidance for TSM for 2021
Parsed Answer: No Info
Ground Truth Answer: The capex guidance for TSM for 2021 is not explicitly stated in the transcript. However

```

```

count = 0
for result in results:
    if result['Parsed Answer'] == 'No Info':
        count += 1

```

```

else:
    print('Question:', result['Question'])
    print('Parsed Answer:', result['Parsed Answer'])
    print('Ground Truth Answer:', result['Ground Truth Answer'])
print(count)

```

```

↳ Ground Truth Answer: The name of the company being discussed in the transcript is TSMC (Taiwan Semiconductor)
Question: What is the expected growth rate for TSMC's revenue from 2020 to 2025 in US dollar terms
Parsed Answer: TSMC projects CAGR to increase by 10% to 15% from 2020 to 2025 in USD terms.
Ground Truth Answer: The expected growth rate for TSMC's revenue from 2020 to 2025 in US dollar terms is 10%
Question: What is the company's gross margin for the 2020-Q4 period
Parsed Answer: TSMC's gross margin improved to 54% in Q4 2020.
Ground Truth Answer: I do not know the company's gross margin for the 2020-Q4 period as it is not mentioned
Question: What was the percentage of wafer revenue contributed by 5-nanometer process technology in the fourth quarter of 2020
Parsed Answer: 20%
Ground Truth Answer: The percentage of wafer revenue contributed by 5-nanometer process technology in the fourth quarter of 2020 was 20%.
Question: What is the company's gross margin guidance for the first quarter of 2021
Parsed Answer: TSMC is expecting a gross margin between 50.5% and 52.5% for Q1 2021.
Ground Truth Answer: The company's gross margin guidance for the first quarter of 2021 is 51.5%.
Question: What is the company's current profitability target as a percentage of gross margin
Parsed Answer: 50.5% and 52.5% for Q1 2021
Ground Truth Answer: The company's current profitability target is 50% of gross margin.
Question: What was the gross margin in Q4 2020
Parsed Answer: TSMC's gross margin improved to 54% in Q4 2020.
Ground Truth Answer: I do not know. The transcript does not provide the gross margin for Q4 2020.
Question: What is the current gross margin of TSMC for the 2022-Q3 period
Parsed Answer: 60.4%
Ground Truth Answer: I do not have access to the current gross margin of TSMC for the 2022-Q3 period as it is not mentioned in the transcript.
Question: What is the gross margin for TSMC in Q3 2022
Parsed Answer: 60.4%
Ground Truth Answer: I do not know the gross margin for TSMC in Q3 2022 from the given transcript.
Question: What is the company's gross margin for the 2022-Q3 period
Parsed Answer: 60.4%
Ground Truth Answer: The company's gross margin for the 2022-Q3 period is not mentioned in the transcript.
Question: What was the gross margin for TSMC in the third quarter of 2022
Parsed Answer: 60.4%
Ground Truth Answer: The gross margin for TSMC in the third quarter of 2022 was 60.4%.
Question: What was the gross margin in the third quarter of 2022
Parsed Answer: 60.4%
Ground Truth Answer: The gross margin in the third quarter of 2022 was 60.4%.
Question: What is the expected gross margin for TSMC in the fourth quarter of 2022
Parsed Answer: The expected gross margin for TSMC in the fourth quarter of 2022 is between 59.5% and 61.5%.
Ground Truth Answer: The expected gross margin for TSMC in the fourth quarter of 2022 is between 59.5% and 61.5%.
Question: What was ASML's gross margin for the 2020-Q1 period
Parsed Answer: 45.1%
Ground Truth Answer: I do not know. The transcript does not provide information on ASML's gross margin for the 2020-Q1 period.
Question: What is the current customer demand for ASML's products and what is the expected customer demand for 2022
Parsed Answer: Strong customer demand
Ground Truth Answer: The current customer demand for ASML's products is strong, and the expected customer demand for 2022 is also strong.
Question: What was the net sales for ASML in Q1 2020
Parsed Answer: €2.4 billion
Ground Truth Answer: The net sales for ASML in Q1 2020 were EUR2.4 billion.
Question: What was the net system sales for Logic in Q1 2020, as a percentage of total net system sales?
Parsed Answer: 73%
Ground Truth Answer: The net system sales for Logic in Q1 2020 was 73%, which is 27% of the total net system sales.
Question: What was the gross margin for ASML in Q1 2020
Parsed Answer: 45.1%
Ground Truth Answer: The gross margin for ASML in Q1 2020 was 45.1%.
Question: What was the gross margin for ASML in 2020-Q1
Parsed Answer: 45.1%
Ground Truth Answer: The gross margin for ASML in 2020-Q1 was not mentioned in the transcript.
238

```

```

results_df = pd.DataFrame(results)
results_df

```


	Ticker	Question	Parsed Answer	Ground Truth Answer
0	TSM	What is the expected growth rate of global sma...	No Info	The expected growth rate of global smartphone...
1	TSM	What is the expected penetration rate for 5G s...	No Info	The expected penetration rate for 5G smartpho...
2	TSM	What is the expected revenue growth rate for T...	TSMC projects CAGR to increase by 10% to 15% f...	The expected revenue growth rate for TSMC fro...
3	TSM	What is the expected revenue from TSMC's back-...	No Info	The expected revenue from TSMC's back-end ser...
4	TSM	What is the expected full volume production of...	No Info	The expected full volume production of SoIC i...
...
261	ASML	What is the estimated revenue that ASML will g...	No Info	The estimated revenue that ASML will get in Q...
262	ASML	What was the total revenue for ASML in 2020-Q1	No Info	I do not know the total revenue for ASML in 2...
263	ASML	What was the gross margin for ASML in 2020-Q1	45.1%	The gross margin for ASML in 2020-Q1 was not ...
264	ASML	What was the order intake for ASML in 2020-Q1	No Info	I do not know the exact order intake for ASML...
265	ASML	What was the net income for ASML in Q1 2020	No Info	The net income for ASML in Q1 2020 was EUR391...

266 rows x 4 columns

```
df_earnings_new = filtered_df_ticker.merge(results_df[['Ticker', 'Question', 'Parsed Answer']],
                                          left_on=['question'],
                                          right_on=['Question'],
                                          how='left')
```

```
df_earnings_new_csv = df_earnings_new.copy()
df_earnings_new_csv.drop(columns=['word_count', 'summary', 'summary_word_count', 'clean_summary', 'Ticker', 'Question'],
df_earnings_new_csv
```