

**Note:**

**This file explained the end-to-end implementation RAG pipeline (0 cost)**

## Contents

Data Source and Dataset Description.....	2
Data Preprocessing and Chunking.....	3
Chunks Embedding and Vector Database .....	4
Prompt engineering for Answer Generation .....	6
Discussion and Future Work .....	9

# Data Source and Dataset Description

## Dataset URL (GitHub repository):

<https://github.com/soumyaah/FinRED/?tab=readme-ov-file>

## Dataset Description:

- 4713 different Earning Call Transcripts(ECT) mainly from the 2020 fiscal year
- Each transcript consists of a .txt file that contains the full transcript for the conference call. Having an average of 41666 Characters per page.

## Sample

Hey, good morning everybody. I'm Ross Seymore, the Semiconductor Analyst here at Deutsche Bank and we're ready to get started with the next presentation of the 2020 Deutsche Bank Technology Conference. We're very pleased to have Forrest Norrod, the SVP and GM of the Datacenter and Embedded Business Solutions Group at Advanced Micro Devices. So, good morning Forrest and thanks for joining us. Good morning. Happy to be here. Sorry, I said, good morning. Pleasure to be with you all. Great! Great to see you back in the office as well. So, before we get into your specific portion of the business in the Datacenter side, I wanted to ask you one other question. Overnight, there were some press reports and speculation that about your semi-custom business and while I realize that's not exactly the topic of the day that we're going to be addressing here, I wondered if you could comment a little bit about those reports? Well look, you know as we've commented before, we have a very large ramp coming into the second half of the year, across multiple product lines, really across every business. And you know I can't really comment on any specific product or any specific customer, but demand continues to be extremely strong and the ramp is going about as we expected. Great! Thanks for addressing that in general. So why don't we dive a little bit deeper into your business. So the Datacenter market and the

*Figure 1 Snapshot of ECT about AMD*

## Dataset Choice Motivation

1. Provide concise summary for lengthy and sometimes irrelevant content. RAG pipeline helps to filter out the irrelevant content and extract important chunks only.
2. RAG helps to pinpoint the exact location where the answer to queries located by providing context around it. Facilitate searching in lengthy documents
3. ECT normally contains jargon without sufficient explanation. GEN AI could use its general knowledge to provide context and answer queries.

# Data Preprocessing and Chunking

## Dataset subset selection

- Among the 4713 files, article-3277 has no text message, so it has been removed
- Only first 20 files by the ascending naming order are selected considering computation efficiency.

## Chunking

- Utilizing langchain's RecursiveCharacterTextSplitter, we split the text into chunks of 800 characters or less with around 100 characters overlapped to bridge the context.
- With an average of 42000 characters per file, we can expect  $(42000 / 700) = 60$  chunks per file

```
text_splitter = RecursiveCharacterTextSplitter(chunk_size = 800,
chunk_overlap = 100)
docs = text_splitter.create_documents(texts)
```

### Caveat:

The docs are created from a list of texts where each element contain one whole ECT. Some might be worried that the resulting chunks might contain small sections at the start of first document followed by the small sections at the end of the next document. This could be a problem since they are independent document and provide different context.

To verify this, I figured out the chunks situating close to the end of first document and start of second document. The red-highlighted chunk are obviously shorter than the chunks before and after it, indicating the function is optimally designed to avoid mixing of documents.

Note also each chunk is lesser than 800 characters, all chunks end with complete word instead of being Truncated to avoid losing context.

"not production yet but this simple demonstration that we had at CES last year gives you a glimpse that the future is coming. And so, we're also working with a number of other robotic companies, as well as universities and other strategic partners. We will be -- the anchor tenants of a new research facility at the University of Michigan, opening this fall, called the Ford Motor Company robotics building on the University of Michigan campus. This is a really exciting partnership we have with the University of Michigan. We will use that facility and our partnership with the University of Michigan to help us accelerate the development of next-gen technologies like advanced robotics, drone technologies, bipedal robots, rolling robots. So all of this is coming and we're right in the middle of" (797 Length),

"bipedal robots, rolling robots. So all of this is coming and we're right in the middle of that mix, working in the far in parallel to working in the now and the near to deliver products today, as well as develop the possibilities for tomorrow. So thanks for that question. Alright. With that, I think we'll wrap it up here. Thank you very much to Ken and the whole team at Ford. I think it was a great conversation and thanks everyone for attending. Thank you." (458 Length)

"Ladies and gentlemen, thank you for standing by and welcome to the ERYTECH Business Update and Financial Highlights for the Second Quarter 2020 Conference Call. At this time, all participants are in a listen-only mode. [Operator Instructions] I would now like to hand the conference call over to your speaker today, Gil Beyen, Chief Executive Officer. Please go ahead, sir. Thank you. Good afternoon, good morning [Foreign Language]. Hoping you're all well and safe and thanks for joining us for our earnings call for the second quarter of 2020 and the first half year. We announced our business and financial update yesterday evening. You should be able to access the press release and our earnings presentation on the Investors page of our website under webcast and link to slide show or via the" (796 Length)

Table 1 Chunks Linking the First and Second Documents

# Chunks Embedding and Vector Database

## Embedding Model Source:

all-MiniLM-L6-v2 from hugging face

## 4 ways to embed the data:

1. CLS token's embeddings
2. Mean pooling with Attention Mask vector
3. Normalized vector of 2<sup>nd</sup> method
4. Use built-in function from Sentence Transformer

## Retrieval Efficiency evaluation

- To test the different ways of embedding the data, we search for 50 contexts relevant to AMD earning call transcripts. 46 chunks are estimated to be relevant in response to the question:

"Can you summarize for me the main content in the earning call transcript for AMD".

- We will be using Recall@50 as to evaluate the model retrieval efficiency

### 1. CLS token's embeddings

```
def get_embeddings(texts):
    inputs = tokenizer(texts, padding=True, truncation=True, return_tensors='pt')
    with torch.no_grad():
        outputs = model(**inputs)
    cls_embeddings = outputs.last_hidden_state[:, 0, :] # Take the [CLS] token's embeddings which is a summary of all the tokens
    return cls_embeddings.numpy()
```

Figure 1 Function to encode chunks using CLS token

$\text{Recall@50} = 4 / 46 \approx 0.087$

### 2. Mean pooling with Attention Mask vector

```
def mean_pooling(model_output, attention_mask):
    token_embeddings = model_output[0] #First element of model_output contains all token embeddings
    input_mask_expanded = attention_mask.unsqueeze(-1).expand(token_embeddings.size()).float()
    return torch.sum(token_embeddings * input_mask_expanded, 1) / torch.clamp(input_mask_expanded.sum(1), min=1e-9)

def embedding_mask(texts):
    encoded_input = tokenizer(texts, padding=True, truncation=True, return_tensors='pt')
    # Compute token embeddings
    with torch.no_grad():
        model_output = model(**encoded_input)
    return mean_pooling(model_output, encoded_input['attention_mask'])
```

Figure 2 Mean Pooling the resulting model output with attention mask vector

$\text{Recall@50} = 7 / 46 \approx 0.1522$

### 3. Normalized vector of 2<sup>nd</sup> method

```
normalized_embeddings_mask = F.normalize(docs_embeddings_mask, p=2, dim=1)
```

Figure 3 Normalize vector from 2nd method

Recall@50 = 11 / 46  $\approx$  0.2391

### 4. Use built-in function from Sentence Transformer

```
st_model = SentenceTransformer('sentence-transformers/all-MiniLM-L6-v2')  
st_embeddings = st_model.encode(chunks)
```

Figure 4 Built-in Function from Hugging Face

Recall@50 = 11 / 46  $\approx$  0.2391

Note that 3<sup>rd</sup> method has the exact 50 chunks context as the 4<sup>th</sup> method, indicating the built-in method is likely having the same structure as the normalized mean pooling vector.

## Vector Database

```
d = doc_embeddings_cls.shape[1]  
vector = faiss.IndexFlatL2(d)  
vector.add(doc_embeddings_cls)
```

Figure 5 FAISS Vector Database is used for embedding storage

- L2 Norm is used as a measure of distance between different embeddings
- FAISS from Meta are used to store embeddings

# Prompt engineering for Answer Generation

## Generative AI model

Qwen/Qwen2.5-72B-Instruct from Hugging Face

<https://huggingface.co/Qwen/Qwen2-72B-Instruct>

### Note:

There're 12 queries being executed in Jupyter Notebook (RAG Implementation.ipynb) to evaluate the effectiveness of the Generative AI model and Retrieval Model. The following documentation only recorded comparison worthy of mentioning.

### Query 1

- The usage of Qwen model is through Free Hugging Face Inference API, this is done through making request to the relevant API URL
- This usage of Inference API resulting in the json object returned always contain the query (Highlighted in yellow) preceding the response. Some data pre-processing is required

20. reflected in the company's cash position at the end of June. Second, and today and this is new. ERYTECH has also announced the implementation of a n At The Market or ATM program, which will allow us at ERYTECH's discretion to issue and sell ordinary shares in the form of ADSs, American Depositary y Shares on the NASDAQ markets through its sales agent, Cowen and Company. That will be sales to eligible investors as a price equal or near to the pre vailing market price on NASDAQ, keeping in mind that the issuance of new shares will also be subject to the same 20% dilution limits as for the ABO con vertible notes. Please note that a new shelf registration statement on Form F-3 was filed with the SEC yesterday, which is nearly a rollover of the she lf registration we already filed last year and

/nAnswer the question based on the above context:/nWhat are the list of companies being mentioned inside the documents?

Based on the provided text, the companies that have been mentioned are:

1. **\*\*Wisgrp\*\*** - Mentioned in the context of the company's website for accessing the second quarter results.
2. **\*\*Pintec\*\*** - The CEO and CFO of Pintec are mentioned as participants in the call.
3. **\*\*thyssenkrupp\*\*** - The company's Q3 numbers are being discussed.

Figure 6 Pieces of response through Hugging Face Inference API, query are always inside the response

### Query 4, 5 and 6

- With the structure of the prompt, advance LLM model will point out if the information to the answer of query is not found in the context
- With prompt as simple as:

"Answer the question based only on the following context:\n{context}\nAnswer the question based on the above context:\n{query}\n"

The response is:

The provided context does not contain any specific information related to the KLA Corporation Fourth Quarter Earning Call Transcript. It appears that the context provided is from different earnings call transcripts from Teekay Corporation, Williams Industrial Services Group, and Unique Fabricating, but not from KLA Corporation. Therefore, there are no 5 highlights from the KLA Corporation Fourth Quarter Earning Call Transcript available in the given context. If you have specific details or transcripts from the KLA Corporation call, please share them, and I'll be happy to help identify the highlights.

Figure 7 Response from Qwen2.5-Instruct for Query 4

## Query 7 ("What is the Level 2, Level 3 and Level 4 Ford Motor refers to")

- Even Level 2, Level 3, and Level 4 are not explicitly explained in the context of Vehicle Automation, the model is able to infer from the context and provide information utilizing general knowledge of GenAI

The response is:

Answer:

In the context of Ford Motor and autonomous driving, different levels of automation from Level 2 to Level 4 are distinguished by the extent of a vehicle's ability to manage driving tasks and the role of the driver. Here's a breakdown:

### Level 2 (Advanced Driver Assistance Systems - ADAS):

- \*\*Description\*\*: The vehicle can perform some driving tasks, such as steering, accelerating, and braking, but the driver must remain engaged and monitor the driving environment at all times.

- \*\*Example\*\*: Ford's Mustang Mach-E with Co-Pilot 360 and Active Drive Assist. These systems allow for advanced hands-off, eyes-off capabilities on certain highways but still require the driver to be attentive and ready to take control.

### Level 3 (Conditional Automation):

- \*\*Description\*\*: The vehicle can handle most driving tasks in specific conditions, such as on highways. However, the driver must be ready to take control if the system cannot handle a situation.

- \*\*Key Challenge\*\*: The major challenge is ensuring that the driver can be reliably alerted and can regain control of the vehicle when needed.

- \*\*Example\*\*: Ford is working on developing a Level 3 solution that will allow drivers to be hands-off, eyes-off, and mind-off, meaning they can engage in other activities like reading a book, but they must be able to take over if the system encounters a situation it cannot handle.

Figure 8 Query 7 Response

## Query 8 ("It took AMD 6 quarters to grow from 5% to 10% market share. How do the representatives from AMD think the slope of share gain curve will be in the future?")

- By asking a very specific question, the GenAI model is able to produce a coherent response by linking different chunks using the overlapping part, indicating the intelligence of generation model.

1. industry was your analyst meeting back in early March and no such number was given at that time. But I did want to just get your thoughts on the market share side of things. It took you about six quarters or a year and a half to go from the 5% to the 10% market share. How do you think the slope of that share gain curve looks going forward? Would you be able to double the share again in the same amount of time? Is it more a much fairer way to look at it that adding the incremental five points of share rather than thinking of it as a doubling what happened over that time? Would it happen faster or slower and why? Yes, so you know I think I don't want to get into the mode of you know setting new benchmarks along the way with fine granularity. I think I've always said about share looking

Figure 9a Chunk 1

15. new benchmarks along the way with fine granularity. I think I've always said about share looking forward is, you know we certainly aspire over time to exceed our historical high water mark, which was you know the 26%, 27% unit share and certainly over time I think we've got the roadmap to do so. But we are not going to put a slope on that line or put further intermediate milestones. What we are focused on though is taking the momentum that we built with the ecosystem and our partners and customers with the first generation of EPYC, Naples and now the second generation and continuing to build on that. So Naples for us you know was giving, quite frankly was giving the first 5%. It was getting to having people use AMD back into the market, was having

Figure 910b Chunk 15

10. giving the first 5%. It was getting to having people used to AMD back into the market, was having customers believe that we could deliver a high performance, a component, whether we were doing what we set, and to that end Rome was hugely important, the second generation EPYC processors that we introduced last August, it was hugely important. Not just because it's a great product and you know really the 64 cores offers literally double the performance of our principal competitors offerings at that point in time, so just about the same. Not only offering literally double the performance, the other critical part of Rome is it's exactly what we said we were going to do, and that predictability of execution gives customers confidence that they can invest in putting AMD into their

Figure 9c Chunk 10

- When being prompted a question that required a long response (15 chunks). The retrieval model will lost context and not necessarily retrieving content that are consecutive linked together

The response is:  
 Answer:  
 The representatives from AMD are not willing to set new benchmarks with fine granularity or provide a slope for the share gain curve. They do aspire over time to exceed their historical high water mark, which was around 26% to 27% unit share, and they believe they have the roadmap to achieve this. However, they are primarily focused on leveraging the momentum they have built with their ecosystem, partners, and customers through the first and second generations of EPYC, and continuing to build on that. They aren't giving a specific timeframe or intermediate milestones for future share gains.

*Figure 10 Query 8 response (Only info from 3 chunks are used even 25 chunks are retrieved)*

## Query 9 & 10 ("Can you summarize for me the main content in the earning call transcript for AMD")

- GenAI produce similar results utilizing the context retrieved from vector database and the context from the full ECT text

4. **Customer Confidence:** AMD's consistent execution and high-performance offerings have given customers the confidence to invest in AMD infrastructure.

5. **Competitive Position:** Despite the challenges, AMD remains focused on its strategy and assumes strong competition. The company's 7nm node ramp has been successful, and any hiccups from competitors do not change its strategic approach.

6. **GPU Market:** AMD sees GPUs as a critical component of growth in the datacenter and is committed to making a significant impact in this market.

7. **Future Aspirations:** The company is driven to continue its aggressive growth and is not satisfied with slowing down the rate of market share gains.

8. **Industry Trends:** The pandemic has had mixed impacts, with work-from-home and cloud demand providing a boost, while new workloads have faced challenges due to uncertainty.

Overall, the call emphasized AMD's strong position in the market, its strategic focus, and its commitment to continued growth and innovation.

*Figure 11a Query 9 Response using the 50 chunks retrieved from vector database*

### Manufacturing and Technology

- **7 Nanometer Node:** AMD has successfully ramped up 7 nanometer production, which has been critical for performance and cost efficiency. The company remains focused on maintaining its manufacturing lead and ensuring sufficient supply to meet demand.
- **Chiplet Technology:** The use of chiplet technology has been pivotal in optimizing performance and yield, allowing AMD to access advanced technology nodes more efficiently.

### Competitive Landscape

- **Pricing:** AMD anticipates a highly competitive environment but focuses on delivering superior performance to win on total cost of ownership (TCO) at the system level.
- **ARM-Based CPUs:** While there has been interest in ARM in the server market, AMD remains focused on high-performance x86 solutions. Migrating workloads from x86 to ARM is a significant barrier, and AMD believes its strong performance roadmap will ensure competitiveness.

### GPU Market

- **Aspirations:** AMD sees GPUs as a critical driver of growth in the datacenter, with key areas including VDI (Remote Rendering), traditional HPC, and machine learning. The company is focused on open-source software frameworks and improving developer ease of use to drive adoption.

### Conclusion

- **Future Outlook:** AMD is confident in its ability to continue growing its datacenter business, driven by a strong product roadmap, expanding customer base, and focus on delivering superior performance and TCO.

*Figure 121b Query 10 Response using full text from the ECT*

## Query 10, 11 & 12("Can you summarize for me the main content in the earning call transcript for AMD")



- Using larger chunks tend to have lose context and resulting in worse performance

The response is:

Answer:

The provided passage does not contain a complete earnings call transcript for AMD, but it does include a segment where an executive is answering questions about AMD's strategic direction and market opportunities. Key points include:

1. **Interest in ARM in Data Centers**: The executive acknowledges long-standing interest in ARM architecture for data centers but emphasizes AMD's focus on providing high-performance CPU and GPU solutions. They believe this focus will keep x86 as the dominant architecture in the core market.
2. **GPU as a Growth Driver**:
  - **VDI and Remote Rendering**: AMD sees significant potential in virtual desktop infrastructure (VDI) and remote rendering, especially with the increase in work-from-home trends post-COVID.
  - **HPC**: They are focusing on traditional high-performance computing (HPC) applications such as scientific and medical research.
  - **Machine Learning and AI**: AMD views machine learning and artificial intelligence as critical areas and is working on open-source software frameworks to facilitate development in these domains.
3. **Strategic Approach**:
  - **Software and Coherence**: AMD emphasizes the importance of software and the need to make GPU technology more accessible and user-friendly. They are also integrating high-performance coherent interconnects between CPUs and GPUs to enhance performance.

*Figure 13a Query 11 Response (Using chunk of 4000 character with 100 overlap) 10 chunks for each ECT*

The response is:

Answer:

The provided transcript segments do not contain information from an AMD earnings call. The provided text includes segments from earnings calls for Pintec Technology Holdings Limited, MSG Networks, Unique Fabricating, and Teekay Corporation. If you have a specific part of an AMD earnings call you wish to be summarized, please provide that segment. Otherwise, I can summarize one of the provided companies' earnings calls. Let me know how you would like to proceed.

If you would like a summary of one of the companies mentioned, please specify which company's earnings call you are interested in. For example:

- Pintec Technology Holdings Limited
- MSG Networks
- Unique Fabricating
- Teekay Corporation

I will be happy to summarize the content for you.

*Figure 14b Query 12 Response (Using chunk of 20000 character with 100 overlap) 2 chunks for each ECT*

## Discussion and Future Work

1. A general rule of thumb is to use at least twice as much the estimated chunks required to obtain a decent result for the query
2. For ECT, it's mostly recommended to keep chunks under 2000 with at least 50 overlapping character. Since chunks with larger size tends to lose context and make proper retrieval much more difficult
3. The current model implemented has its bottleneck lies in the retrieval model as its recall rate are relatively low. It could be mostly due to the poor embedding model selection
4. One solution to the problem in query 8 of not able to answer a question that span over large amounts of chunks (15 chunks) is to:
  - a. Employed an hybrid retrieval model, where the retrieved vector can be rerank where chunks that are in sequence are to extracted out together.
  - b. An alternative approach: If identify the question as a response that required context span over huge text (Using GenAI to identify), then after retrieving, implement a system that would fetch all the consecutive chunks to form a larger chunk before sending the prompt to GenAI.

