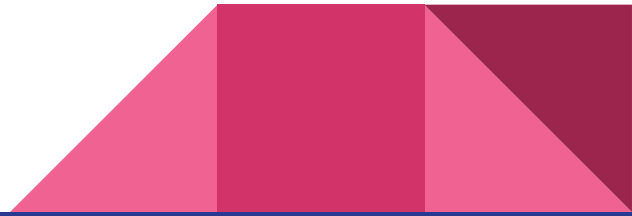
Research Analysis

Research Topics

1. Preparation of Data
2. Market Analysis
3. RAG Model

Methodology For Preparation of Data

1. Web Scraping
2. Translation
3. Convert to Markdown file
4. Creating Chunks for Table Data



Web Scraping

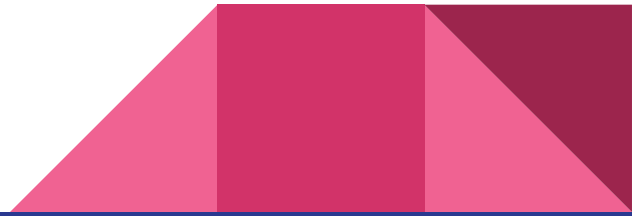
- KAP datas’ are chosen to be investigated for web.

  https://www.kap.org.tr/

- Three firms are chosen for testing the scraping algorithm which are Koç, Turkcell and Akbank.
- Accessing several documents and downloading them in excel format.

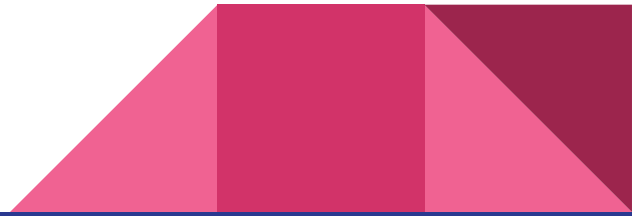
Several Problems in Downloaded Files

- Unfortunately, the KAP download file is an excel file that contains HTML. It is hard to summarize HTML's with LLM.
- The files contain tables which is hard for LLM to summarize. Because a lot of information is lost during summarization
- The language is also Turkish. English is better for LLM.



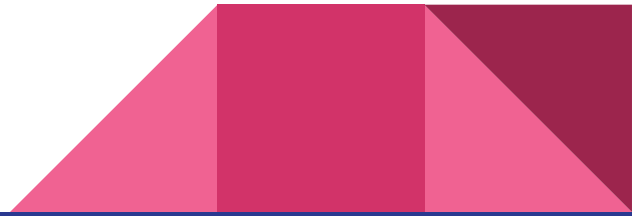
Translation

- For translating purposes, we used Llama 3.1-7B as local LLM. The responses were not satisfied our criteria.
- We decided to use GPT4o as our LLM. The responses are better but it is expensive and requires Internet Connection.
- Each HTML element that contains text is translated individually then replaced with the actual text



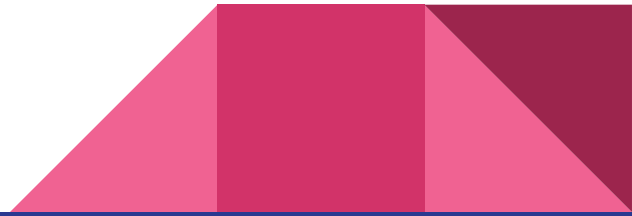
Convert to Markdown file

- LLM understand Markdown file better.
- This process yields writing a automatization code manually. Downside is impossible to cover all possibilities.



Creating Chunks for Table Data

- In order to create knowledge graph, summarization for table data is needed.
- Since table data is big, sending all at once cause information loss
- Chunking methods are used



Chunk Algorithms

- Fixed-Length Chunking
- Semantic or Sentence-Based Chunking
- Sliding Window Technique
- Hierarchical Chunking

Semantic Chunking

- Similarity coefficient is chosen to be 0.8 for better results.
- Ex.

Coefficient = 0.8

```
Chunk 3:
| Dönen Varlıklar | 605.974.596 | 982.090 | 1.712.378 | 2.197.689 |
| Duran Varlıklar | 414.578.711 | 600.504 | 1.146.587 | 1.389.200 |
| Varlıklar | 1.020.553.307 | 1.582.594 | 2.858.965 | 3.586.889 |

Chunk 4:
| Kısa Vadeli Yükümlülükler | 709.676.931 | 1.145.655 | 1.865.813 | 2.403.
| Uzun Vadeli Yükümlülükler | 189.771.314 | 189.741 | 318.880 | 388.901 |
```

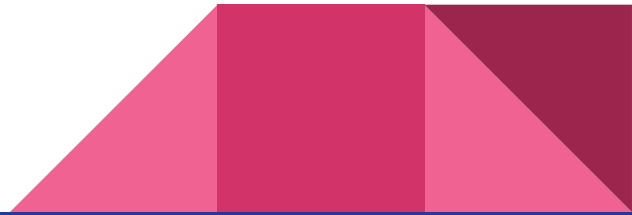
Coefficient = 0.7

```
Chunk 3:
| Dönen Varlıklar | 605.974.596 | 982.090 | 1.712.378 | 2.197.689 |
| Duran Varlıklar | 414.578.711 | 600.504 | 1.146.587 | 1.389.200 |
| Varlıklar | 1.020.553.307 | 1.582.594 | 2.858.965 | 3.586.889 |
| Kısa Vadeli Yükümlülükler | 709.676.931 | 1.145.655 | 1.865.813 | 2.403.087 |
| Uzun Vadeli Yükümlülükler | 189.771.314 | 189.741 | 318.880 | 388.901 |
```

Market Analysis

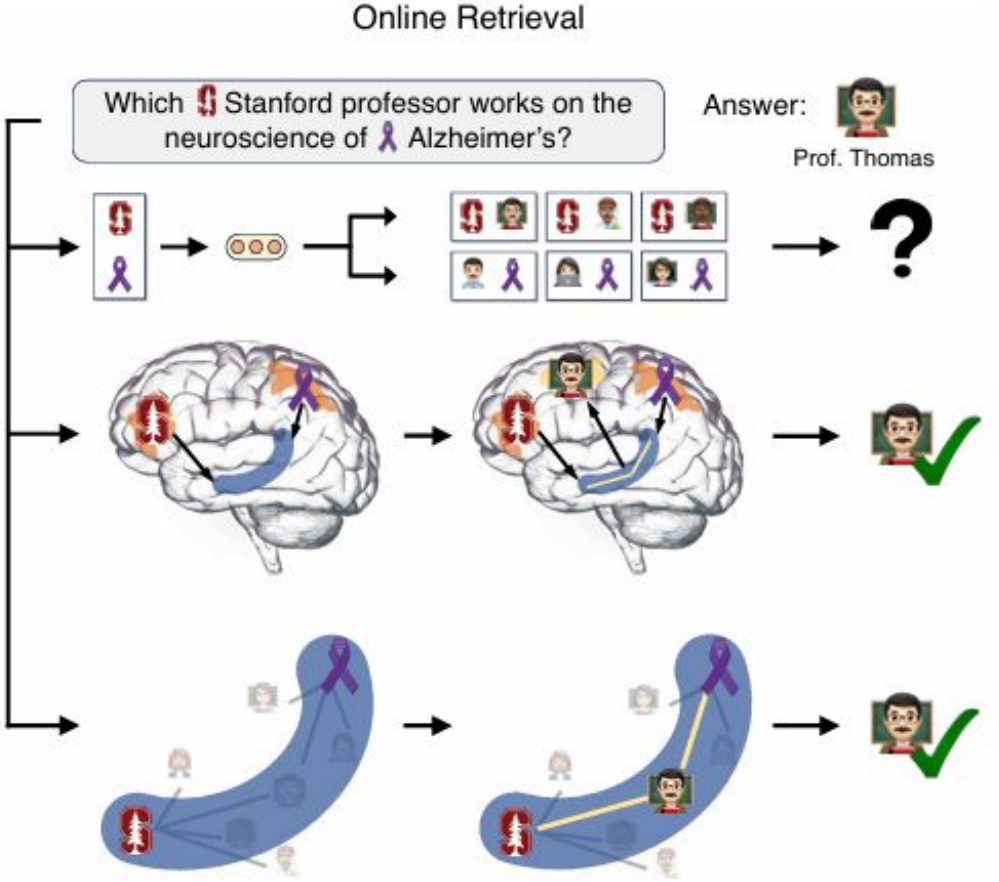
- **FinChat**: has notifications (?, to what extent we don't know) but lacks multi-hop and comparative question capability in chatbot
- **AI Ticker Chat**: no notifications, meeting analysis over transcript etc.
- **Uptrends.ai**: has notifications but context is discretized, not over language like ours, which can cover more continuous contexts

None of them are specialized to Borsa Istanbul Companies, some companies even don’t exist in their systems.



Multi-hop Questions

We should be able to accurately answer questions that require information from separate parts of the knowledge base. Example →



Comparative Questions

Ex: What are the top 5 companies with the highest profits in the last year?

Needs different chunks from different places, and also the ability to do computations on them.

Current RAG systems are not well on these tasks

We designed a new RAG architecture to solve these problems