Raw Data

Data Preparation For Finetune 1

We will have three category data for our Llama fine-tuning

- Data from FinGPT (5)
  - https://huggingface.co/datasets/FinGPT/fingpt-sentiment-train
  - https://huggingface.co/datasets/FinGPT/fingpt-headline-cls
  - https://huggingface.co/datasets/FinGPT/fingpt-ner-cls
  - https://huggingface.co/datasets/FinGPT/fingpt-finred
  - https://huggingface.co/datasets/FinGPT/fingpt-finred-cls
- Other source from hugging face
  - ChanceFocus/flare-cfa
  - jan-hq/finqa_bench_stealth-finance-v3
- **Data Extracted from our own** (Question generated by gpt-3.5-turbo 🤯)
  - CFA textbook, --- cover profolio management, equity investment, derivatives, financial statement, fixed income.
  - Macro research paper (From Bloomberg) :
    - 500 query
  - Equity research report (From Bloomberg)
    - 10 companies-> 300 query
  - Blooberg daily morning news : total 500 news ---> convert to 200 query
  - the final version of our own data can be found in the `data/Finetune1_data_pre/final_data`

Raw data for Finetune 2

This part is all about data prepare for finetune 2

Data source

- US stock market dow 30
  - data source: https://huggingface.co/datasets/FinGPT/fingpt-forecaster-dow30-202305-202405
- H share stock market (HSI top 10, HSI Tech Top 10)
  - **data created by ourselves**
  - stock price : akshare
  - news: bings, akshare
  - fundemental : akshare
  - the final version of the data can find in `data/Finetune2_data_pre/final_output`