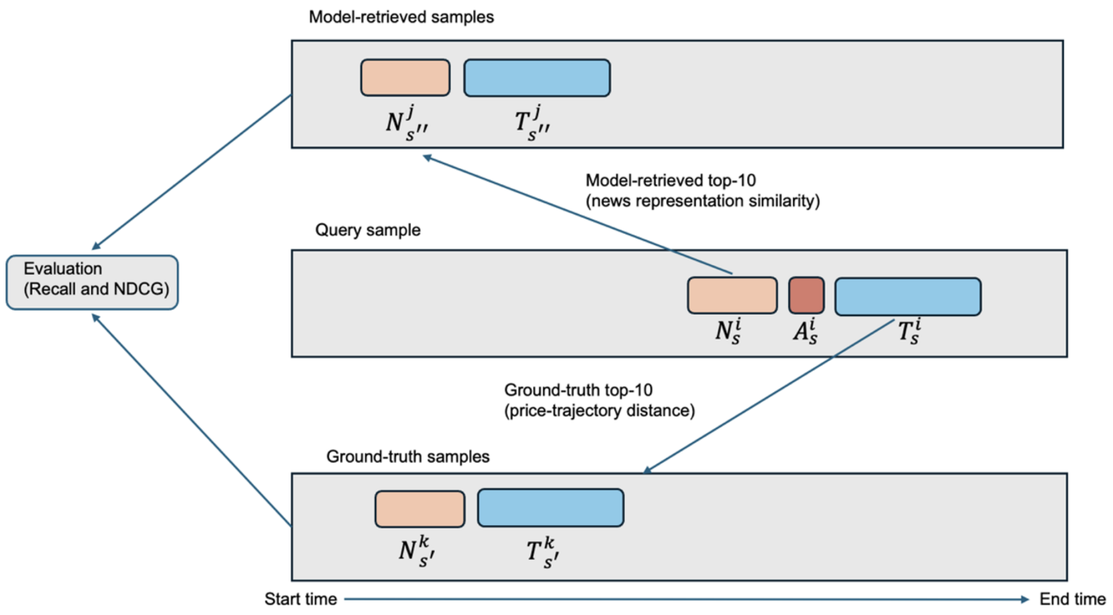
Retrieving Financial News with Forward-looking Time Series Analogy

SMU SINGAPORE MANAGEMENT UNIVERSITY
School of Computing and Information Systems

Li Du
Supervisor: Professor Ma Yunshan

Research Problem

This research aims to retrieve financial news with similar subsequent stock price time series.



Dataset Construction & Preprocessing

We downloaded both the price and news data of Nasdaq-100 between Year 2021 and 2025. News data comes from Massive and Alpha Vantage API. However, both news datasets have serious problems. Price data is from Yahoo Finance.

Massive Problems	Alpha Vantage Problems
Duplicate news Multi-ticker assignment Truncated news Non-news content	Multi-ticker assignment Time uneven distribution Time misalignment

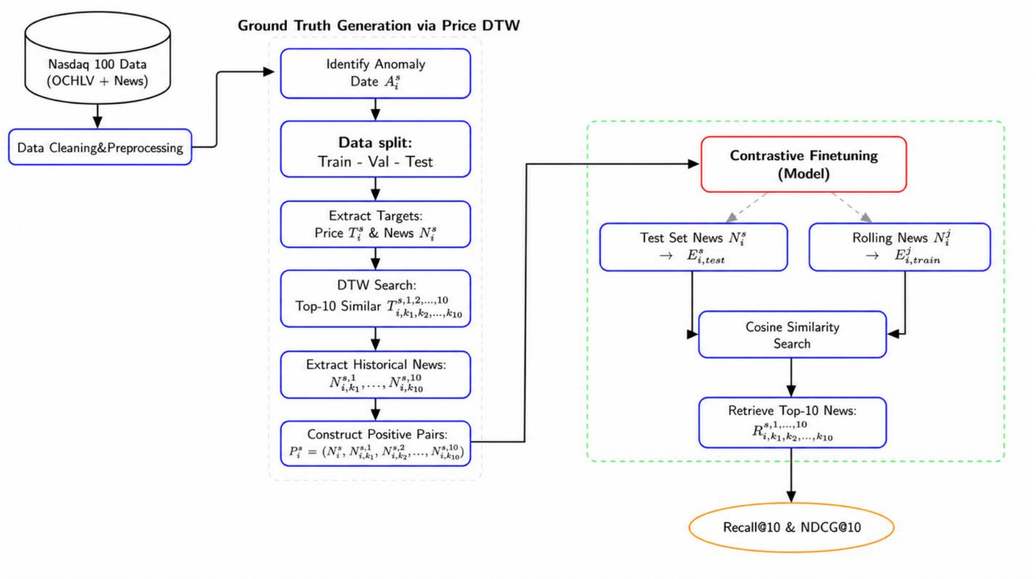
Datasets deep cleaning

Rule-based screening → LLM group-level labelling → Rule-based ticker-level label mapping → Keep / Downweight / Drop / Review

24,000 scrape-error or empty-summary rows dropped
Multi-ticker rate: >50% → 10.8%
52,000+ junk or non-news rows removed
60% of news in 2025 reduced to less than 37.5%
30%+ of publication dates realigned to trading days
399,490 merged rows → 93,154 keep-labelled rows

Methodology

Price-trajectory similarity defines relevance, while textual similarity is learned to retrieve the corresponding historical news contexts.



1. Anomaly Identification
5d news + anomaly + 10-d trajectory

2. DTW Ground Truth
Top-10 market-wide most similar time-series

3. Training
Agg / Sum / Pool → BGE

4. Evaluation
Cosine retrieval → Recall / nDCG

Two-Level News Representation: Daily News → 5-Day Window

Aggregation	**LLM Summarization**	Pooling: Max/Avg/Attention
Daily: Concatenate articles 5D: Concatenate daily text	Daily: Summarize articles 5D: Summarize daily summary	Daily: Pool article embedding 5D: Pool daily vectors

Anomaly Detection

$$z_{s,t} = \frac{r_{s,t} - \mu_s}{\sigma_s}, \quad |z_{s,t}| > 2$$

DTW Ground Truth

$$G_i^{10} = \mathrm{TopK}_j\left[-DTW(T_i, T_j)\right]$$

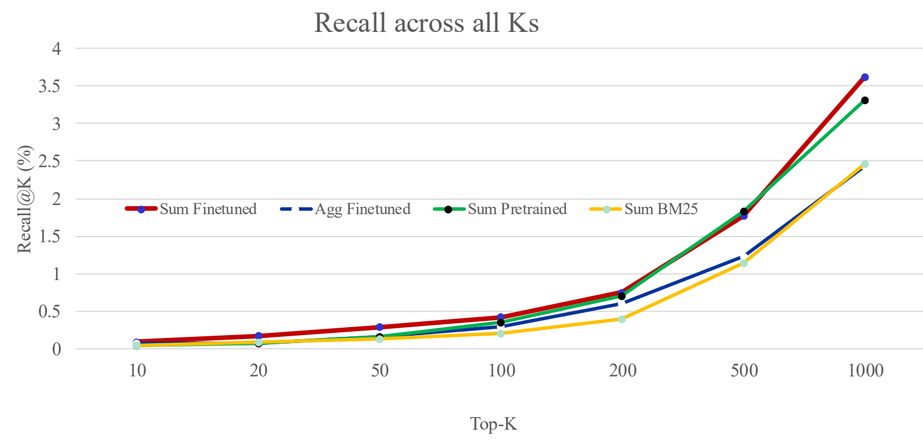
Contrastive Learning Loss

$$\mathcal{L}_i = -\log \frac{\exp(\mathrm{sim}(u_i, v_i^+)/\tau)}{\sum_{j \in B_i} \exp(\mathrm{sim}(u_i, v_j)/\tau)}$$

DTW-matched news contexts are treated as positives; other in-batch contexts serve as negatives.

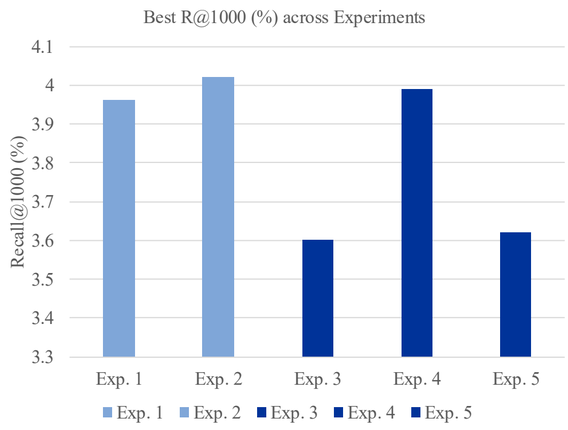
Results

LLM summarization was the best performing model, but absolute retrieval performance Recall@1000 remained below 5%, revealing a fundamental paradigm challenge.



Sum Finetuned outperforms Agg Finetuned at K=1000
The performance gap widens as K increases

Performance stayed within a narrow range across experiments.



Cleaning and context fusion did not boost performance

* *Valid-news filtering in Exp. 5.3 reached the highest R@1000 of 4.50%*

Exp.1	Original data+Simple prompt
Exp.2	Original data+Simple prompt with market fusion
Exp.3	Clean data+Improved prompt
Exp.4	Clean data+Simple prompt with market fusion
Exp.5	Clean data+Improved prompt with market fusion

Hypothesis Summary

H1:SUM>AGG	H2:Pooling Attn>Max/Avg	H3:Cleaning+ Context
Mostly (4/5) supported	**Mixed evidence**	**Not supported**

Embedding space diagnostics

Pretrained BGE	Finetuned BGE
Vector Mean Sim: 0.51	Vector Mean Sim: 0.97
>0.95 Sim Ratio: 0.04%	>0.95 Sim Ratio: 97%

Fine-tuning compressed most news embeddings into a nearly indistinguishable region.

* Sim refers to Cosine similarity between embeddings

Discussion & Future Work

Why is this task so difficult?

1. The time window being too long
Compared with other studies that retrieve news information within one day, our research experiments on a 5-day news window, which is much harder because the information to be summarized increased by 5 times. Also the time-series window is 10-day long, so the impact of news could decay fast as days go by

2. The embedding model might not be the best option
Text embedding models like BGE are pretrained on semantic similarity, which might not be suitable for capturing the structured relationship like "news+price reactions"

3. Strict ground-truth evaluation
Recall/NDCG only count DTW matches as correct, therefore other related cases, even though still highly similar to the query, must be counted as misses and dropped.

Future work

1, Simplify the task by reducing the window length for both news and price time series.

2, Experiment with other frameworks that could capture a structured relationship such as a "cause-and-effect" relationship rather than only semantic similarity

3, Read the two stage LLM summarization and compare it with human written summarization. Also explore other news representation methods that could potentially outperform the three types: Agg, Sum, Pool

4, Use DTW-distance-weighted score in the loss instead of treating all DTW top-10 matches equally in the contrastive learning

5, Use a cleaner and more valid dataset like FNSPID, or generate an artificial dataset just to test the right learning frameworks to learn the patterns we want the model to learn

Contributions

- On the task level, we propose a new framework that is designed to retrieve financial news with similar subsequent stock price pattern;
- On the method level, we design and implement a series of experiments to investigate the proposed task.
- We show that fine-tuned LLM summarization is the best performing representation at larger retrieval depths, while Recall@1000 remains below 5%, revealing fundamental limitations in the current embedding framework.