MURA-Finance: Multi-Hop Reasoning with Augmented Context for Implicit Financial Sentiment Analysis

Team: Team 1
Team Members: Long Nguyen, Quynh Nguyen, Johnny

Glossary of Abbreviations

- **MURA-Finance: Mu**lti-hop **R**easoning with **A**ugmented Context for **Finance**.
- **THOR: T**hree-**ho**p **R**easoning.
- **CoT: C**hain-**o**f-**T**hought.
- **RAG: R**etrieval-**A**ugmented **G**eneration.
- **LLM: L**arge **L**anguage **M**odel.

1. Article

“Reasoning Implicit Sentiment with Chain-of-Thought Prompting” - Hao et al., 2023

Hao et al. (2023) introduced the THOR (Three-hop Reasoning) framework to address the challenge of "implicit sentiment" - opinions expressed without explicit sentiment words. Instead of direct classification, the model reasons sequentially using a Chain-of-Thought (CoT) pipeline consisting of 3 hops:

Inferring Aspect → Inferring Opinion → Inferring Polarity

Their results show improved accuracy and explainability compared to single-step prompting approaches.

2. DOI / Link

https://arxiv.org/pdf/2305.11255

3. Objective

Building upon the THOR framework (Hao et al., 2023), which utilizes a 3-hop reasoning process for general implicit sentiment, this project introduces **MURA-Finance** as a specialized extension to the **financial domain**, where sentiment is often indirect and closely tied to market interpretation. While the base paper focuses on general product aspects, we expand the chain of thought into a **5-hop domain-specific pipeline** designed to navigate the linguistic complexities of financial markets - such as hedging and indirect warnings - and translate inferred sentiment into actionable market signals .

Stakeholders & Beneficiaries

This framework serves financial analysts, investors, traders, currency strategists, and algorithmic trading systems by providing transparent, explainable reasoning for market sentiment.

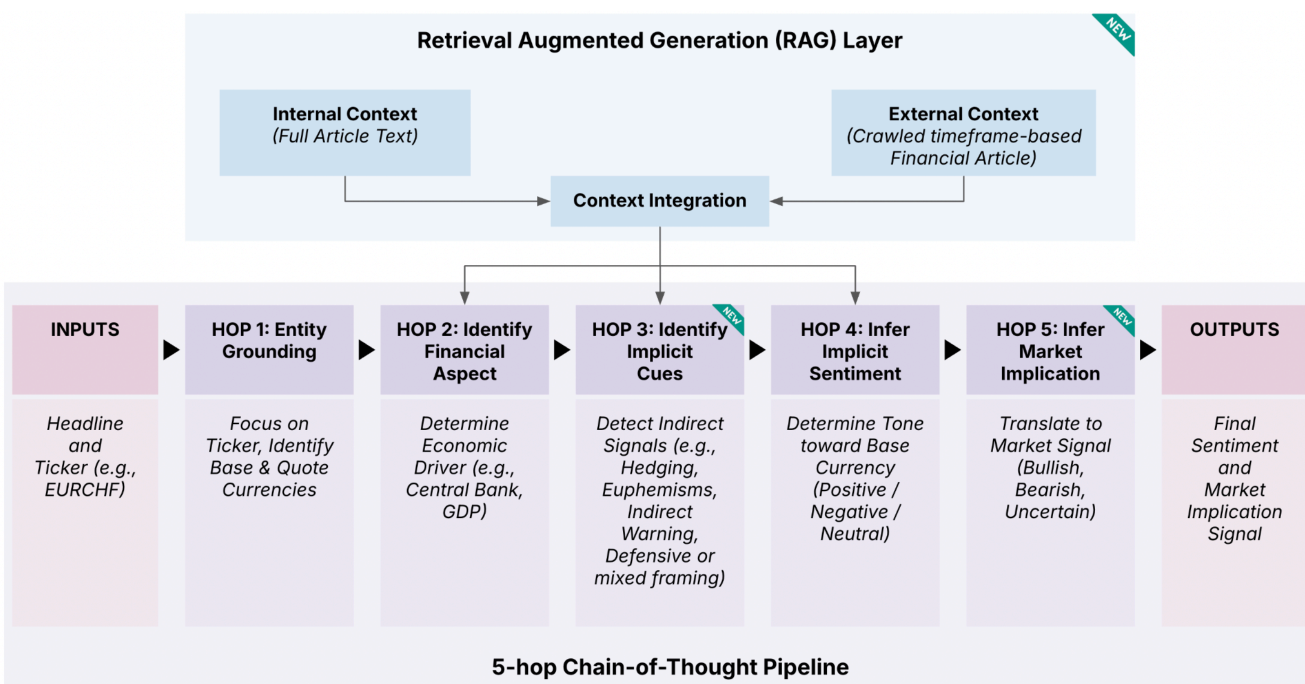
Use Case

Given a ticker like EURUSD and a subtle headline such as *"The ECB remains cautious despite easing inflation headwinds"*, the model identifies the "cautious" and "headwinds" as Implicit Cues. It then reasons that while the tone is guarded, the reduction in headwinds implies a strengthening Euro relative to the Dollar, resulting in an implicit sentiment of Positive and a Bullish market implication for the base currency (EUR).

4. Methods

4.1. Core Methodological Approach

To realize the objective, we will employ a 5-hop CoT pipeline enhanced by Retrieval-Augmented Generation (RAG).



Detailed 5-Hop CoT Pipeline

- **Hop 1 - Entity Grounding:** The model uses the provided “ticker” from the dataset (e.g., EURCHF) to focus its analysis. It identifies which currency is the "base" and which is the "quote" within the context of the title.
- **Hop 2 - Identify Financial Aspect:** Identifies the economic driver mentioned (e.g., Central Bank determination, GDP data, or Inflation).
- **Hop 3 - Identify Implicit Cues (Addition):** Detect indirect linguistic signals commonly used in financial communication:
    - Hedging (e.g., “may”, “expects”, “remains cautious”)
    - Euphemisms (e.g., “challenging environment”)
    - Indirect warnings (e.g., “headwinds”, “pressure persists”)
    - Defensive or mixed framing
- **Hop 4 - Infer Implicit Sentiment:** Determines the tone toward the base currency (Positive/Negative/Neutral).
    - Positive: implies improvement, strength, or recovery for the base currency.
    - Negative: implies deterioration, weakness, or pressure on the base currency.
    - Neutral: presents factual data or balanced views with no clear directional bias for the base currency.
- **Hop 5 - Infer Market Implication (Addition):** Translates sentiment into Bullish, Bearish, or Uncertain signals.
    - Bullish: Expectation that the exchange rate will rise.
    - Bearish: Expectation that the exchange rate will fall.
    - Uncertain: Evidence is contradictory, hedged, or insufficient to determine a clear market direction.

Note on Architectural Flexibility: *While we propose an initial 5-hop structure, during the experimental phase, we will iteratively test and optimize the reasoning chain. If empirical results or error analysis suggest that fewer or more steps provide better logical consistency and accuracy, the architecture will be adjusted accordingly.*

Retrieval-Augmented Generation (RAG) Layer (Addition)

We will retrieve minimal supplementary context to improve the model's understanding of short headlines:

- **Internal Context:** Content from the "text" column of the dataset (full article content).
- **External Context:** 3-10 related financial articles via timestamp-based news crawling (financial-news-multisource dataset, Yahoo Finance, Reuters, etc.).

4.2. Evaluation and Strategy:

Metrics

Accuracy, Precision, Recall, and F1-score

Ablation Strategy

We will perform a comparative analysis across three distinct levels of complexity, using FinBERT as our primary specialized baseline:

- **Level 1: Baselines (FinBERT & Single-shot Classification):**
  - FinBERT: We will use the pre-calculated scores in the primary dataset to establish a performance floor for specialized transformer models.
  - Single-shot LLM: Direct, 1-hop sentiment classification without intermediate reasoning steps.
- **Level 2: 5-hop Reasoning (No RAG):**
  - Testing the performance of the domain-specific 5-hop chain using only the headline and ticker as input.
- **Level 3: Full MURA-Finance (5-hop + RAG):**
  - The complete pipeline where the reasoning hops are supported by retrieved context from the full article text and external multisource news.

Error Analysis

A manual review of 100 discordant samples will be performed to check if the "reasoning chain" broke at the Cue Detection (Hop 3) or Aspect (Hop 2) phase.

5. Review of Prior Work

Hao et al., 2023

- **Title:** Reasoning Implicit Sentiment with Chain-of-Thought Prompting
- **Link:** https://arxiv.org/pdf/2305.11255
- **Achievement:** Established the THOR framework, which uses sequential Chain-of-Thought steps to solve implicit sentiment in general domains.
- **Limitation:** It is not optimized for financial linguistics.

Fatouros et al., 2023

- **Title:** Transforming Sentiment Analysis in the Financial Domain with ChatGPT
- **Link:** https://arxiv.org/pdf/2308.07935
- **Achievement:** Evaluated Large Language Models (LLMs) on FX news headlines, setting a benchmark for zero-shot and few-shot classification in the currency market.
- **Limitation:** Focused on 1-hop direct classification, which misses the explainable reasoning required to decode implicit market "cues" like central bank hawkishness.

Kangtong et al., 2024

- **Title:** Fine-Tuning Gemma-7B for Enhanced Sentiment Analysis of Financial News Headlines
- **Link:** https://arxiv.org/pdf/2406.13626

- **Achievement:** Demonstrated that fine-tuned open-source 7B models can outperform general-purpose LLMs on domain-specific financial sentiment tasks.
- **Limitation:** Focuses on literal sentiment classification without exploring the multi-hop reasoning chain or the specific market implications (Bullish/Bearish) of the findings.

6. Originality of the Project

This project introduces a novel, explainable, and market-oriented approach to implicit financial sentiment analysis by combining multi-hop Chain-of-Thought reasoning with domain-specific modeling and contextual augmentation. These capabilities are not addressed by existing financial sentiment methods.

The main original contributions are:

(1) **First Financial-Domain Multi-Hop CoT Framework**
We expand the THOR framework from three to five reasoning hops, introducing a dedicated hop for implicit financial cue detection, capturing hedging, euphemisms, and indirect signals prevalent in financial text.

(2) **Market Implication Inference**
Beyond sentiment polarity, we introduce a new step that infers market implications (bullish, bearish, uncertain), providing actionable interpretations aligned with financial decision-making.

(3) **Retrieval-Augmented Context Integration**
We incorporate a lightweight RAG layer to enrich short financial headlines with relevant context and systematically evaluate its contribution to reasoning accuracy.

(4) **Explainable and Interpretable Outputs**
The structured, hop-wise outputs enable transparent analysis of how implicit cues lead to sentiment and market interpretation, improving trust and usability over black-box classifiers.

7. Individual Contribution

<table>
<tr><th>Task</th><th>Sub-task</th><th>Owner</th><th>Clickup Ticket</th></tr>
<tr><td rowspan="2">Task 1: Data Preparation</td><td>Collect & clean datasets</td><td>Long</td><td rowspan="2"><link></td></tr>
<tr><td>Convert labels to sentiment + implication</td><td>Long</td></tr>
<tr><td rowspan="2">Task 2: 5-Hop CoT Pipeline</td><td>Design 5-hop prompts</td><td>Johnny</td><td rowspan="2"><link></td></tr>
<tr><td>Implement reasoning pipeline</td><td>Johnny</td></tr>
</table>

Task	Sub-task	Owner	Clickup Ticket
Task 3: Model Benchmarking	Benchmark against: Fine-Tuning Gemma-7B for Enhanced Sentiment Analysis of Financial News Headlines (as per *Kangtong et al., 2024*)	Long	<link>
	Benchmark against: Transforming Sentiment Analysis in the Financial Domain with ChatGPT (*Georgios et al., 2023*)	Quynh	<link>
Task 4: RAG Layer	New data crawling for context enhancement	Quynh	<link>
	Build retriever + context generator	Quynh	<link>
Task 5: Experiments & Analysis	Run evaluations + produce report	Long, Johnny, Quynh	<link>
Task 6: Presentation	Slides + visualization	Long, Johnny, Quynh	<link>

8. Data Source and Description

Primary Dataset

Forex Financial News Headline Dataset

- **Source:** Georgios et al. (2023)
- **Link:** https://arxiv.org/abs/2308.07935
- **Size:** 2,291 expert-labeled headlines.
- **Attributes:** published_at, ticker, true_sentiment, title, author, url, source, text, finbert_sentiment, finbert_sent_score
- **Nature:** This dataset provides ground-truth labels for specific FX pairs (e.g., EURUSD, GBPUSD).

RAG Contextual Database

Financial News Multisource

- **Source:** Ferrell (2024)
- **Link:** https://huggingface.co/datasets/Brianferrell787/financial-news-multisource

- **Nature:** A large-scale, multi-source financial news aggregator.
- **Utility:** This dataset serves as the External Knowledge Base for our RAG layer. When the model processes a short headline from the primary dataset, it will query this multisource corpus to retrieve relevant articles and economic reports published around the same timeframe to provide a broader market context.

Secondary Datasets

Financial PhraseBank

- **Link:** https://huggingface.co/datasets/takala/financial_phrasebank
- **Size:** 4,840 sentences categorized by sentiment agreement.
- **Utility:** Used for generalization tests to ensure the model maintains accuracy across broader financial news beyond the FX market.

FiQA / SemEval-2017 Task 5 (subtask 2)

- **Link:** https://github.com/tocab/SemEval2017Task5
- **Size:** ~1,143 news headlines.
- **Utility:** Essential for testing **Hop 2 (Aspect Identification)** and **Entity Grounding**, as this dataset includes fine-grained aspect-based sentiment scores.

9. Computational Details

Hardware: Google Colab Pro, CPU: ≥8 vCPUs, RAM 16 GB

Software: Python 3.x, PyTorch / HuggingFace Transformers, LangChain for multi-hop and RAG pipeline, Jupyter notebooks for experimentation, Vector DB (ChromaDB / FAISS)

Model Hosting/Access: GPT-5 API model, Gemini 2.5 Flash, DeepSeek V3.2

10. References

[1] Y. Hao, J. Wang, W. Hong, and D. Zhang, "Reasoning Implicit Sentiment with Chain-of-Thought Prompting," *arXiv preprint arXiv:2305.11255*, 2023. [Online]. Available: https://arxiv.org/abs/2305.11255

[2] X. Kangtong, "Fine-Tuning Gemma-7B for Enhanced Sentiment Analysis of Financial News Headlines," *arXiv preprint arXiv:2406.13626*, 2024. [Online]. Available: https://arxiv.org/abs/2406.13626

[3] G. Fatouros et al., "Transforming Sentiment Analysis in the Financial Domain with ChatGPT," *arXiv preprint arXiv:2308.07935*, 2023. [Online]. Available: https://arxiv.org/abs/2308.07935

[4] L. Havrlant and V. Takala, "Financial PhraseBank," *HuggingFace Datasets*, 2013. [Online]. Available: https://huggingface.co/datasets/takala/financial_phrasebank

[5] K. Schlegel et al., "SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News," 2017. [Online]. Available: https://github.com/tocab/SemEval2017Task5

[6] B. Ferrell, "Financial News Multisource Dataset," *HuggingFace Datasets*, 2024. [Online]. Available: https://huggingface.co/datasets/Brianferrell787/financial-news-multisource