SCUNLP-1 at the NTCIR-18 FinArg-2 Task : Collaborative Large Language Models for Temporal Classification

Min-Chin Ho, Jheng-Long Wu
Department of Data Science, Soochow University, Taiwan

Introduction

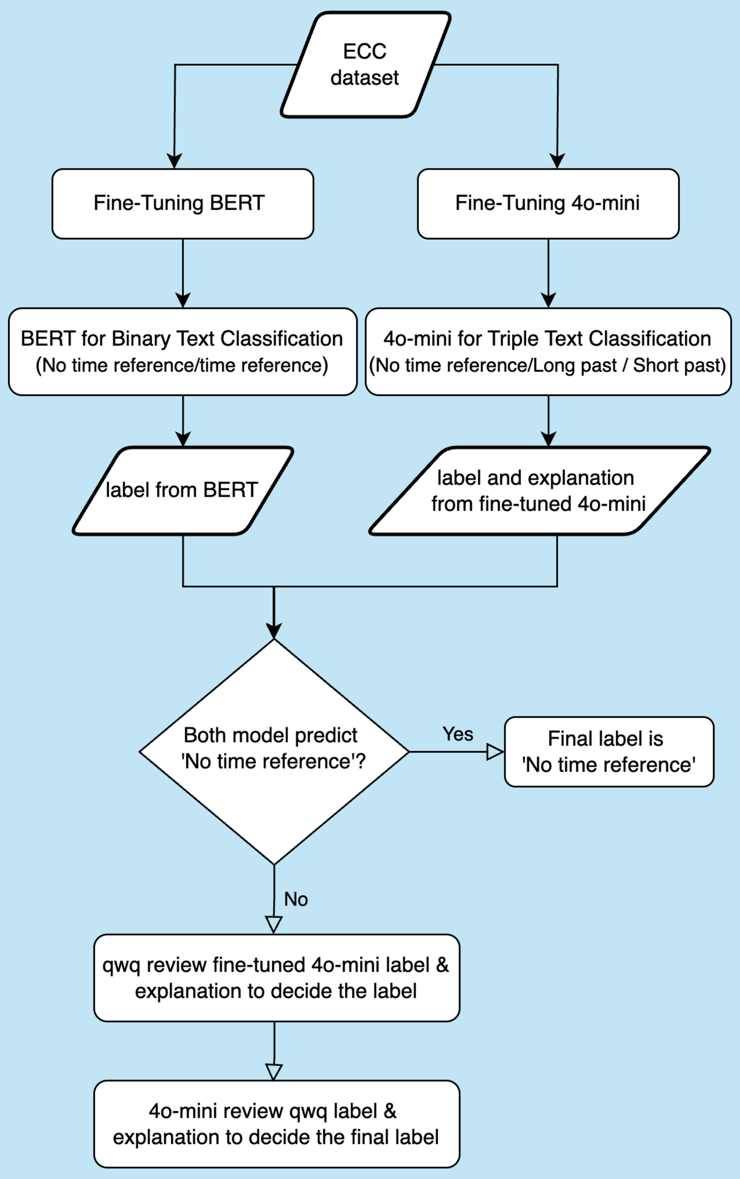
In recent years, understanding temporal information within financial texts has become essential for capturing the dynamics of market events and investor behavior. While prior research has extensively studied sentiment and argument mining, the role of temporal reasoning in financial argumentation remains relatively unexplored. This study addresses this gap by leveraging large language models (LLMs) to detect and analyze temporal references in financial arguments. Through novel collaboration mechanisms and tailored prompt designs, we aim to enhance the accuracy and interpretability of temporal predictions, providing deeper insights into the timing and duration of financial impacts.

Purpose

- Identify temporal references in financial arguments and assess their impact on the expected duration of financial effects.
- Use large language models (LLMs) to perform temporal argument mining in financial texts.
- Apply negotiation-style prompting to guide the model in extracting and reasoning about time-related information.
- Simulate multi-agent or self-reflective dialogues to uncover and evaluate time-sensitive claims.

Method

The flow of our study provides a comprehensive explanation of how large language models (LLMs) assign temporal tags to financial texts. First, we introduce the overall decision-making process. Then, we detail the design of custom prompts for each LLM, enabling them to analyze temporal elements and perform self-reflective reasoning.



Our approach adopts a two-stage framework for temporal argument classification. In the first stage, we use a fine-tuned BERT model to perform binary classification, distinguishing between “Time Reference” and “No Time Reference” content. A fine-tuned o4o-mini model is also used for verification. If both models predict “No Time Reference,” the instance is finalized with that label. Otherwise, it proceeds to the second stage for further analysis.

In the second stage, we address BERT’s limitations in assessing time duration by leveraging the advanced reasoning capabilities of LLMs. Through a negotiation-style framework, multiple LLMs evaluate temporal cues and collaboratively determine the most appropriate temporal label.
To support temporal reasoning, we design a prompt framework that guides LLMs through step-by-step reasoning using Chain of Thought. Each LLM is placed in a multi-turn, debate-style setting, where it reviews the prediction and explanation generated by the previous model. This allows the model to validate or challenge prior conclusions, encouraging deeper reasoning and producing more consistent and interpretable temporal classifications.

first LLM : fine-tuned 4o-mini

```
Task:Given a claim and its supporting premises from an Earnings Conference Call, determine if the argument contains references to past time periods.

The labels should be chosen based on the following guidelines:
Label 0 (No time reference): The claim and premises do not provide explicit or clear time references, or they talk about future actions with no reference to past events.
Label1 (Long past): The claim or premises reference events or trends that occurred more than six months ago (past two quarters).
Label 2 (Short past): The claim or premises reference events that happened within the past six months, such as the last quarter or during the current quarter.

claim:{claim}
premises:{premises}
Current Time Context(Quarter and Year):{quarter} {year}

Steps to follow:
1.Examine the Claim: Check if the claim contains any temporal references (e.g., past events, specific timeframes like 'last quarter', 'this year', etc.).
2.Analyze the Premises: Look for evidence of temporal cues in the premises (e.g.,'this quarter', 'last year', 'earlier this year').
3.Understand the Time Context: Consider the year and quarter mentioned in the data and how the claim relates to the premises.
4.Assign the Correct Label:
-Label 0 if no specific past time is referenced or the focus is on future plans.
-Label 1 if the reference is to long-past events (more than six months ago).
-Label 2 if the reference is to short-past events (within the last six months or current quarter).
```

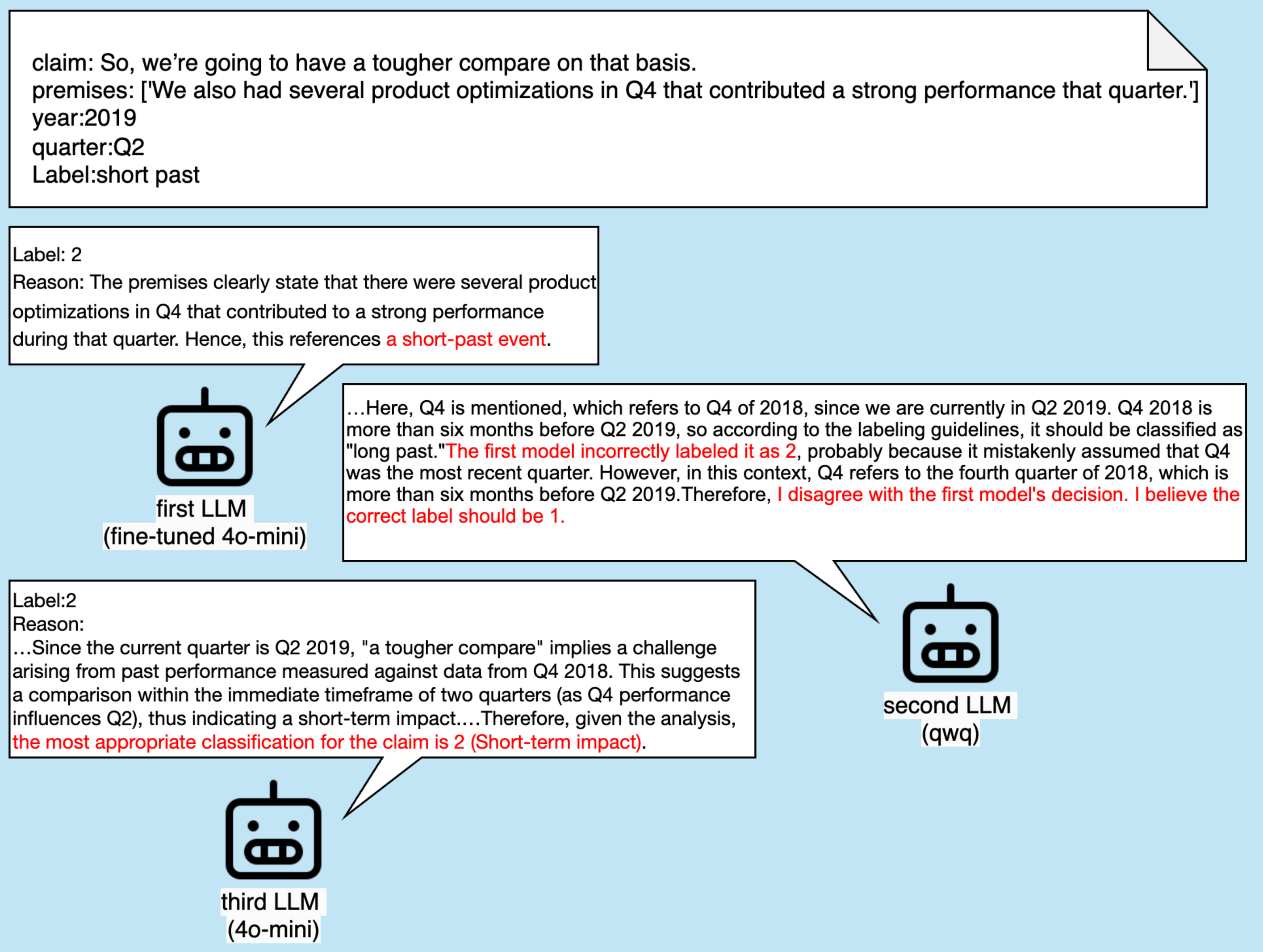
second LLM : qwq

```
Task:Given a claim and its supporting premises from an Earnings Conference Call, determine if the argument contains references to past time periods.
The labels should be chosen based on the following guidelines:

Label 0 (No time reference): The claim and premises do not provide explicit or clear time references, or they talk about future actions with no reference to past events.
Label 1 (Long past): The claim or premises reference events or trends that occurred more than six months ago (past two quarters).
Label 2 (Short past): The claim or premises reference events that happened within the past six months, such as the last quarter or during the current quarter.

claim: {claim}
premises: {premises}
Current Time Context(Quarter and Year): {quarter} {year}

Based on the first model's statement:{the explanation of the first LLM}
Do you agree with the first model's decision that the temporal tag for the input sentence is {the label of the first LLM}?  If not, what would be your suggested label?

Steps to follow:
1.Examine the Claim: Check if the claim contains any temporal references (e.g., past events, specific timeframes like 'last quarter', 'this year', etc.).
2.Analyze the Premises: Look for evidence of temporal cues in the premises (e.g.,'this quarter', 'last year', 'earlier this year').
3.Understand the Time Context: Consider the year and quarter mentioned in the data and how the claim relates to the premises.
4.Assign the Correct Label:
-Label 0 if no specific past time is referenced or the focus is on future plans.
-Label 1 if the reference is to long-past events (more than six months ago).
-Label 2 if the reference is to short-past events (within the last six months or current quarter).
```

third LLM : 4o-mini

```
Task:Given a claim and its supporting premises from an Earnings Conference Call, determine if the argument contains references to past time periods.
The labels should be chosen based on the following guidelines:

Label 0 (No time reference): The claim and premises do not provide explicit or clear time references, or they talk about future actions with no reference to past events.
Label 1 (Long past): The claim or premises reference events or trends that occurred more than six months ago (past two quarters).
Label 2 (Short past): The claim or premises reference events that happened within the past six months, such as the last quarter or during the current quarter.

claim:{claim}
premises:{premises}
Current Time Context(Quarter and Year):{quarter} {year}

Step-by-step reasoning:
Step 1: Identify whether the claim contains any explicit time references.
- If the claim contains a time reference (e.g., mentions "this quarter", "last year", "in 2023"), note the specific timeframe.
If no time references are mentioned, classify as 0 (No time reference) and conclude the analysis.
- Key Improvement: Ensure that the absence of a time reference leads directly to classification as 0, with no further analysis needed.
If the claim lacks specific time indicators, do not attempt to infer or assume time frames, and immediately classify as 0.

 Step 2: If the claim contains a time reference, analyze the premises to determine the impact's duration.
- For claims with explicit time references, analyze the premises to assess whether the event is short-term (within 6 months or 2 quarters) or long-term (over 6 months or 2 quarters).
- Pay attention to any time-related references in the premises, such as “this quarter”, “next year”, “year-over-year” to help assess the temporal scope.
- When the claim explicitly mentions a time period, such as specific quarters or years, ensure to correlate the time period with the event described.
- Key Improvement: If the claim refers to ongoing trends or long-term impacts (e.g., "over the years", "over the next two years"), classify as long-term (1),
even if the premises mention short-term data or specific quarterly details.
- Key Improvement: If the claim involves a specific event happening in the past or upcoming quarters, classify as short-term (2), even if the impact is described as sustained in the premises.
This ensures that claims with clear short-term references are classified as such.

Step 3: Correlate the year and quarter information provided with the claim and premises.
- Cross-check the year and quarter to verify if the events in the claim and premises are tied to the current or past/future quarters.
- Key Improvement: If the timeframe mentioned spans beyond 6 months or 2 quarters, classify as long-term (1). If it is within 6 months or 2 quarters, classify as short-term (2).
- Key Improvement: If the claim refers to future events or trends over a longer period (e.g., “over the next two years”), it should be classified as long-term (1),
even if the premises only describe specific events in the near future. This ensures accurate classification for claims about future trends.
- Key Improvement: Verify if events occurred within the current or prior quarters before making a decision.
This ensures that if there are references to past or future quarters, those should be cross-checked with the claim to verify if they match the expected timeframe of events.

Step 4: After analyzing both the claim and premises, determine the correct temporal classification.
- Based on the analysis of explicit time references, the duration of the event in the premises, and any correlations with year/quarter information,
classify the claim as either 0 (No time reference), 1 (Long-term impact), or 2 (Short-term impact).
- Key Improvement: Make sure to focus on both the claim’s and the premises' temporal references, avoiding assumptions when references are unclear.
If a claim discusses both short-term and long-term effects, ensure to prioritize the overall context of the claim—whether it is describing a trend or a specific event.

Step 5: Final Check Before Classification
- Reassess the claim and premises to confirm their temporal scope.
- Claims referencing long-term trends (e.g., \"year-over-year\") are typically classified as long-term (1), even if premises include short-term events.
- Short-term events tied to specific quarters without broader context should be classified as short-term (2).
- Focus on whether the claim describes ongoing trends or isolated short-term occurrences.

Refer to other models' decisions, and decide the final label:
first LLM's Decision: {the label and explanation of the first LLM}
second LLM's Decision: {the label and explanation of the second LLM}
```

Submission & Result



- Official Results

The official test results reveal that Run 3 achieved the highest overall performance, suggesting that our final prompt design—featuring both step-by-step reasoning and access to prior model outputs—was the most effective. This underscores the importance of prompt clarity and information flow in optimizing LLM-based classification.

RUN	LLM 1	LLM 2	LLM 3	Final Decision (LLMs + BERT)
RUN 1	71.43%	61.9%	65.48%	66.67%
RUN 2	71.43%	67.86%	52.38%	58.33%
RUN 3	71.43%	66.67%	63.1%	67.86%

- Predictions of Each LLM and Final Decision

The per-model performance shows that LLM1 consistently performed well across all runs, confirming its reliability as the first-stage reasoner. Variability in LLM2 and LLM3 outcomes reflects sensitivity to input structure and the quality of prior explanations. The final decision performance fluctuates depending on how effectively downstream models leverage earlier reasoning.

RUN	Micro F1	Macro F1
SCUNLP-1_ECC_3	67.86%	64.94%
SCUNLP-1_ECC_1	66.67%	63.06%
SCUNLP-1_ECC_2	58.33%	52.07%

- Prediction Results Across LLM Stages

Stage-wise breakdown highlights the internal dynamics of the multi-turn reasoning process. While most samples are correctly classified by LLM1, about 28% require further deliberation. Notably, LLM3 occasionally overrides correct predictions from LLM2, indicating that later-stage models may prioritize their own logic over prior answers. This reinforces the need to better align reasoning consistency and trust across models.

Stage	RUN1	RUN2	RUN3
Correct in LLM1	71.43%	71.43%	71.43%
Transition to LLM2	28.57%	28.57%	28.57%
Transition to LLM3	38.10%	32.14%	33.33%
LLM2 Correct, LLM3 Rejected	4.76%	26.20%	10.71%

Conclusion

This study proposes a two-stage temporal labeling framework combining BERT with multi-turn large language model (LLM) collaboration for classifying time-related arguments in financial texts. Our findings demonstrate that step-by-step reasoning and inter-model reference to prior predictions and explanations can enhance both classification accuracy and interpretability.

While there remains room for improvement in model design, reasoning flow, and computational efficiency, our results highlight the potential of collaborative LLMs in tackling complex temporal understanding tasks. Future work may further explore interaction strategies between models, prompt engineering, or variations in model scale to strengthen temporal reasoning performance.