Fine-Tuning and Prompt-Based Methods for Temporal Reasoning in Multilingual Financial Texts

Bor-Jen Chen [1][0009-0002-5239-8308], Wen-Hsin Hsiao[1][0009-0003-4714-2049], Hsin-Ting Lu [1] and Min-Yuh Day[1*] [0000-0001-6213-5646]

[1] Graduate Institute of Information Management, National Taipei University, New Taipei City, Taiwan
s711336103@gm.ntpu.edu.tw, s711336107@gm.ntpu.edu.tw, s711436119@gm.ntpu.edu.tw, myday@gm.ntpu.edu.tw

Abstract. Temporal reasoning in financial texts is essential for understanding event timing and claim validity, especially in earnings conference calls and social media discussions. While transformer-based models have advanced natural language processing, the comparative performance of fine-tuned encoder models and prompt-based decoder models in multilingual temporal classification remains underexplored. This study systematically compares model types, model sizes, and prompting strategies across two tasks: detecting temporal references in English texts and assessing claim validity in Chinese posts. Encoder models such as RoBERTa and BERT and decoder models such as GPT-4o, Mistral, and Gemma are evaluated using fine-tuning and few-shot prompting approaches. Results show that fine-tuned encoder models achieve consistently strong performance across both English and Chinese datasets. Mid-sized prompt-based decoder models also perform competitively under well-designed prompts, offering a practical alternative when fine-tuning is not feasible. In addition, decoder models are more robust to class imbalance, as reflected by smaller gaps between Micro-F1 and Macro-F1 scores. However, decoder models perform less effectively on Chinese tasks, indicating the need for language-specific adaptation. These findings provide practical guidance for selecting models and designing prompts for financial natural language processing under resource constraints.

Keywords: Financial NLP, Temporal Reasoning, Fine-Tuning, Prompt-Based Learning, Large Language Models (LLMs)

1 Introduction

Recent advancements in natural language processing (NLP) and the rapid development of large language models (LLMs) have created new opportunities for understanding financial texts, especially in extracting temporal cues from earnings conference calls and social media discussions. These sources often contain time-sensitive insights that are essential for evaluating market trends, analyzing financial arguments, and supporting timely decision-making. However, temporal reasoning in financial

texts remains challenging due to the implicit nature of temporal expressions, varied linguistic patterns, and domain-specific complexities across languages. For example, determining whether a company's forecast refers to the upcoming quarter or a longer-term projection often requires interpreting vague expressions like "soon" or "in the near future. These phrases may not have fixed meanings and must be interpreted in context, sometimes relying on document metadata or implicit event ordering. Such complexity makes it difficult for models to extract accurate temporal information without strong contextual understanding.

To address these challenges, two main modeling paradigms have emerged: fine-tuned encoder-based models and prompt-based decoder models. Encoders such as RoBERTa and BERT typically require task-specific fine-tuning and perform well with sufficient labeled data. In contrast, decoder models like GPT-4o and Mistral are increasingly used in in-context learning, offering flexibility in zero-shot and few-shot settings without further training. While both approaches have shown promise, few studies systematically compare their effectiveness across languages, model sizes, and prompt configurations.

This study aims to fill the gap by investigating encoder-based fine-tuning and decoder-based prompting methods for temporal classification in financial texts. Two datasets are used to represent different language and domain contexts: the English-based Earnings Conference Call (ECC) dataset and a Chinese Social Media dataset. These settings enable a comprehensive evaluation of how different model types, sizes, and prompting strategies affect performance across scenarios. Specifically, this study addresses three research questions: (1) How do fine-tuned encoder models compare with decoder-based models using prompt-based inference? (2) Can small or mid-sized language models perform competitively with large models in in-context settings? (3) How does language difference influence model performance across tasks?

The contributions of this study are threefold. First, it presents a comparative evaluation of encoder-based fine-tuning and decoder-based prompting across languages and model sizes. Second, it highlights the viability of mid-sized decoder models, such as Mistral-24B and Gemma-27B, which achieved performance comparable to larger models or fine-tuned encoders. Third, it sheds light on prompt design considerations and language-specific challenges in temporal reasoning, providing practical insights for both researchers and practitioners working in multilingual financial NLP.

The remainder of this paper is structured as follows: Chapter II reviews related work on temporal reasoning and model architecture. Chapter III describes the research methodology, including datasets, models, and evaluation setup. Chapter IV presents experimental results and analysis. Chapter V concludes the study with key findings, implications, and future directions.

2 Related Work

2.1 Temporal Reasoning in Finance

Temporal reasoning is essential in financial NLP, particularly when identifying how statements relate to specific time periods or have lasting effects. A labeled dataset from earnings conference calls was created, with each argument annotated based on its temporal reference: short past, long past, or no time mention. Statements with historical grounding tended to be viewed as more persuasive and reliable, suggesting that temporal framing plays a role in shaping perceived argument quality.[1].

Another work focused on the lasting impact of financial news by introducing impact duration awareness during pre-finetuning. News content was categorized based on how long its effects persist, such as short, medium, or long term. This approach emphasized that effective temporal reasoning involves not only detecting time expressions but also modeling how long financial information remains relevant [2].

The challenge of evaluating forecasting skill from text has also been examined by aligning natural language predictions with actual outcomes. Findings indicate that forward-looking statements vary in predictive value, and capturing this variation is essential for accurate temporal inference. [3].

These studies collectively demonstrate that temporal reasoning in finance involves more than surface-level time detection. It requires classifying temporal references, modeling duration, and evaluating the predictive quality of statements, which are all vital for downstream tasks such as forecasting and financial claim analysis.

2.2 Fine-Tuning Strategies for Encoder Models

Fine-tuning pre-trained encoder models such as BERT has become a standard strategy in financial NLP tasks [4]. Adjusting hyperparameters like batch size, learning rate, and training epochs can lead to significant improvements in [5]. In financial domains, task-specific fine-tuning has been shown to enhance performance on domain-sensitive inputs. For instance, fine-tuned FinBERT has been used to classify sentiment in forward-looking statements, where input structure and domain knowledge jointly contribute to better accuracy [6].

Recent advances have introduced more robust fine-tuning methods. One approach involves contrastive adversarial training, which creates semantically similar adversarial examples to improve generalization and resistance to input variations. This method can be applied to any encoder model and improves performance under domain shifts [7]. Additionally, combining BERT with sequential models such as LSTM has proven effective in financial risk prediction tasks, enabling the capture of temporal dependencies while retaining semantic richness from pre-trained encoders [8].

2.3 Prompting Methods with Decoder Models

Prompting methods with decoder-based large language models (LLMs) have become a practical alternative to fine-tuning, especially in tasks involving few-shot learning. Unlike encoder-based models, decoder-only architectures like GPT excel at processing natural language prompts to perform classification, reasoning, and generation. A foundational overview of prompt-based learning categorizes methods such as zero-shot, few-shot, and chain-of-thought prompting, and emphasizes the importance of aligning prompt templates with the model's pretraining objectives [9].

In the financial domain, few-shot prompting has shown promising results. One study evaluated ChatGPT's performance on sentiment, stance, and topic classification tasks using three-shot examples embedded in the prompt. The results demonstrated that even without task-specific fine-tuning, ChatGPT achieved competitive performance across various financial classification settings, highlighting the capability of decoder models to generalize from limited in-context demonstrations [10].

Further expanding the prompting landscape, a comprehensive catalog of prompt patterns was introduced to guide systematic prompt engineering with ChatGPT. This catalog includes strategies like step-by-step reasoning, role prompting, and example-based prompting, each tailored to enhance LLM responses under different task conditions. The catalog emphasizes that the structure and clarity of prompts play a critical role in maximizing model performance [11].

These studies underscore the effectiveness of prompting in decoder-based models and demonstrate how carefully crafted prompt structures can serve as a viable, low-resource approach to financial text classification.

2.4 Model Scaling and Small Language Models

Model scaling plays a critical role in financial NLP, especially when balancing performance and computational cost. Although large language models (LLMs) such as GPT-4 have demonstrated strong performance in various tasks, their deployment requires significant computational resources and infrastructure support [12]. In contrast, recent research has emphasized the growing potential of small language models (SLMs), particularly when enhanced by optimization techniques such as quantization and instruction tuning. These models can achieve competitive results while maintaining lower memory usage and reduced latency [13].

The trade-off between model size and efficiency becomes more evident in real-world scenarios. One study examined serving architectures for SLMs and demonstrated how to achieve Pareto-optimal throughput with minimal accuracy loss, offering practical strategies for low-latency deployment [14]. In addition, comparative analysis has shown that model scale alone does not guarantee better task performance, and that effective adaptation and alignment with task objectives are also essential [15].

With appropriate training and prompt strategies, SLMs offer a cost-effective and efficient alternative, especially in environments where computational resources are limited.

3 Methodology

3.1 Research Framework

This study aims to compare fine-tuning and prompt-based learning strategies for temporal reasoning in financial texts using both encoder-based and decoder-based language models. Temporal reasoning refers to the task of determining the time relevance or validity of a statement, which is crucial for interpreting financial claims in context. The overall research framework is illustrated in Figure 1.

This study focuses on two temporal classification tasks from the FinArg-2 dataset: the English-based ECC Temporal Reference task and the Chinese-based Social Media Claim Validity task. These datasets undergo preprocessing steps, including text cleaning, standardization, and data augmentation (only for the ECC task to balance label distribution).

We adopt two training strategies based on the model type: The encoder-based models (RoBERTa-base, DistilBERT-base, BERT-Chinese, and DistilBERT-multilingual) are fine-tuned using supervised learning on their respective datasets. The decoder-based models (GPT-4o, Mistral-24B, Gemma3-27B, Mistral-8B and Gemma2-9B) are evaluated using prompt-based learning with zero-shot, three-shot, and six-shot settings.

All models are evaluated using Micro-F1 and Macro-F1 scores on a fixed validation set to ensure consistency and comparability across models, training strategies, and tasks.

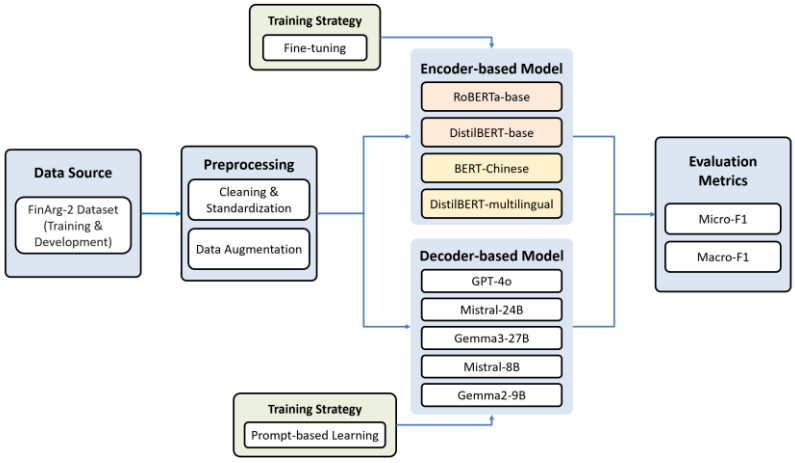

Fig. 1. Overall research framework for temporal reasoning with fine-tuned encoders and prompt-based decoders.

3.2 Datasets and Preprocessing

This study utilizes two sub-datasets from the FinArg-2 competition: the Earnings Conference Call (ECC) dataset for English temporal reference classification, and the Social Media dataset for Chinese temporal validity assessment. Each dataset corre-

sponds to a distinct language and task structure, requiring separate preprocessing procedures.

ECC Temporal Reference Dataset (English). This dataset contains claim-premise pairs extracted from earnings call transcripts, along with time metadata (year and quarter). The goal is to classify each instance into one of three categories: 0 = No time reference, 1 = Long past, 2 = Short past.

To address label imbalance in the 601-sample training set, data augmentation was performed using GPT-4o mini to generate semantically equivalent rephrased samples. A controlled prompt ensured the preservation of financial terminology and factual consistency. This yielded a balanced training set of 903 samples, while the 150-sample validation set remained unaltered to reflect real-world distributions. A 10% sample of the augmented data was manually verified.

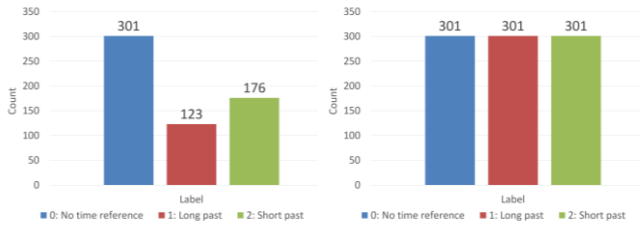

Fig. 2. Label distribution diagram before and after data argumentation on ECC Dataset.

Social Media Claim Validity Dataset (Chinese). This dataset includes Chinese investor comments labeled with claim validity duration: "Longer than 1 week", "Within 1 week", or "Unsure".

The training set consists of 6,132 samples, with notable label imbalance (70% for "Longer than 1 week"). The development set includes 876 samples with similar distribution. No augmentation was performed, and the original imbalance was preserved to evaluate model robustness under real-world data conditions.

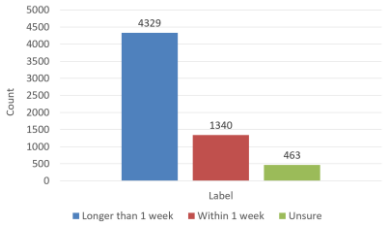

Fig. 3. Label distribution diagram on Social Media Dataset.

3.3 Model Selections

This study compares two categories of models: encoder-based models fine-tuned on labeled data and decoder-based models evaluated using prompt-based inference. All

models were applied to the same two task-specific datasets: the English-language ECC dataset and the Chinese-language Social Media dataset. The selected models vary in size and architecture, enabling a comparative analysis of their performance on financial temporal reasoning tasks.

Encoder-Based Models. To establish strong supervised learning baselines, transformer-based encoder models were fine-tuned on the corresponding datasets. The selected models are listed in Table 1.

Table 1. Encoder-Based Transformer Models Fine-Tuned on task-specific datasets.

Model Name	Size	Dataset	Notes
RoBERTa-base	125M	ECC (English)	Strong encoder baseline pretrained on large English corpora
DistilBERT-base	66M	ECC (English)	Lightweight, efficient BERT variant, suitable for faster fine-tuning.
BERT-Chinese	102M	Social Media (Chinese)	Pretrained on Chinese corpora, suitable for monolingual tasks.
DistilBERT-multilingual	134M	Social Media (Chinese)	Compact multilingual model, adaptable to cross-lingual tasks.

Decoder-Based Models. Decoder-based language models were evaluated using prompt-based inference without weight updates. Each model was tested under zero-shot, three-shot, and six-shot settings. The selected models are shown in Table 2.

Table 2. Decoder-Based Language Models Used for Prompt-Only Inference.

Model Name	Size	Source	Notes
GPT-4o	100B+	OpenAI	Powerful closed-weight model, used as high-end baseline
Mistral Small-3.1-24B Instruct	24B	Mistral	Medium-scale instruction-tuned open model
Gemma-3 27B-it-qat	27B	Google	Quantization-aware instruction-tuned model for efficient inference
Ministral-8B	8B	Mistral	Small decoder model, open-weight and fast inference
Gemma2 9B	9B	Google	Latest lightweight model for general-purpose language tasks

These decoder models represent a range of large language model scales. GPT-4o serves as a high-end baseline, while the open-weight Mistral and Gemma variants allow for performance evaluation under constrained computational budgets.

3.4 Training and Inference Procedure

Encoder Fine-Tuning. Transformer-based encoder models were fine-tuned using supervised classification on both datasets. Inputs were preprocessed using each model's tokenizer by adding special tokens ([CLS], [SEP]), truncating or padding to a

fixed length (128 or 256 tokens), and generating input IDs and attention masks. The models were trained with class labels for 3 to 6 epochs using the AdamW optimizer and cross-entropy loss. Early stopping was based on validation Micro-F1 and Macro-F1 scores. To improve generalization, gradient clipping, weight decay (0.01), and dropout were applied. A grid search over learning rate, batch size, sequence length, and epochs was used to determine the best configuration for each model. Hyperparameter ranges are listed in Table 3.

Table 3. Fine-Tuning Hyperparameter Settings.

Hyperparameter	Values
Learning Rate	1e−5, 1.5e−5, 3e−5, 5e−5
Max Length	128, 256
Batch Size	16, 32, 64, 128
Epochs	3, 4, 5, 6

Decoder Inference Settings. This study evaluates decoder-based language models using a prompt-only inference approach without any parameter updates. All models are tested under three in-context learning configurations: zero-shot, three-shot, and six-shot. This setup enables analysis of how prompt length and the number of demonstrations influence model performance in financial temporal reasoning tasks

Each prompt consists of the following components:

(1) A concise task instruction that clearly describes the prediction goal.
(2) A label definition list with explicit class mappings.
(3) A variable number of in-context examples, depending on the shot setting. For instance, the three-shot setting includes one labeled example per class, while the six-shot setting includes two per class. These examples are randomly sampled from training data previously correctly predicted by the same model.
(4) The test input, typically ending with a label selection cue, such as: " Label:"

All decoder-based models used the same in-context examples for each task. These examples were randomly selected from training data that had been correctly predicted by all models, ensuring consistent quality. This setup helps isolate performance differences caused by model architecture rather than example variability.

3.5 Evaluation Metrics

To evaluate model performance, this study adopts two commonly used classification metrics: Micro-F1 and Macro-F1, which together provide a comprehensive view of prediction quality across both balanced and imbalanced datasets.

Micro-F1 aggregates the contributions of all classes by computing precision and recall across all instances before calculating the F1-score. This metric gives equal weight to each individual prediction, making it particularly suitable for datasets with imbalanced class distributions, where the majority class may dominate the results.

Macro-F1 first computes the F1-score independently for each class, then takes the unweighted average across all classes. This approach ensures that performance on minority classes is fairly represented, regardless of class size.

These metrics were chosen to reflect both overall accuracy (Micro-F1) and the model's ability to handle class imbalance (Macro-F1), especially important given the skewed label distributions in both datasets. By analyzing both scores, we evaluate not only the model's general predictive capability but also its robustness across categories with varying frequencies.

4 Experimental Results and Analysis

4.1 Encoder-Based Model Performance

Table 4. Encoder-Based Models Performance.

Dataset	Model	Micro-F1	Macro-F1
ECC	RoBERTa-base	**69.05%**	67.06%
ECC	DistilBERT-base	65.48%	62.44%
Social Media	BERT-Chinese	**72.83%**	53.40%
Social Media	DistilBERT-multilingual	69.98%	53.50%

Table 4 presents the performance of encoder-based models on both tasks. On the ECC dataset, RoBERTa-base outperformed DistilBERT-base, achieving a Micro-F1 of 69.05% and a Macro-F1 of 67.06%, indicating stronger capability in capturing temporal cues in English financial texts.

For the Social Media dataset, BERT-Chinese achieved the highest Micro-F1 (72.83%), while DistilBERT-multilingual scored slightly lower (69.98%). However, both models had similarly low Macro-F1 scores (around 53%), suggesting challenges in handling underrepresented classes such as "Unsure."

Overall, fine-tuned encoder models serve as strong baselines. However, the gap between Micro-F1 and Macro-F1 indicates the impact of class imbalance, especially in more imbalanced datasets.

4.2 Decoder-Based Model Performance

Table 5. Prompt-Based Decoder Model Performance on the ECC Dataset (English).

Model	Size	Prompt Setting	Micro-F1	Macro-F1
GPT-4o	Large	0-shot	65.48%	58.57%
GPT-4o	Large	3-shot	**67.86%**	62.36%
GPT-4o	Large	6-shot	66.67%	62.33%
Mistral-24B	Medium	0-shot	58.33%	48.09%
Mistral-24B	Medium	3-shot	64.29%	59.79%

Mistral-24B	Medium	6-shot	**69.05%**	64.43%
Gemma-27B	Medium	0-shot	66.67%	60.45%
Gemma-27B	Medium	3-shot	67.86%	64.50%
Gemma-27B	Medium	6-shot	67.86%	61.36%
Mistral-8B	Small	0-shot	61.90%	54.69%
Mistral-8B	Small	3-shot	**64.29%**	52.89%
Mistral-8B	Small	6-shot	59.52%	42.80%
Gemma2-9B	Small	0-shot	48.81%	47.33%
Gemma2-9B	Small	3-shot	55.95%	51.78%
Gemma2-9B	Small	6-shot	60.71%	57.81%

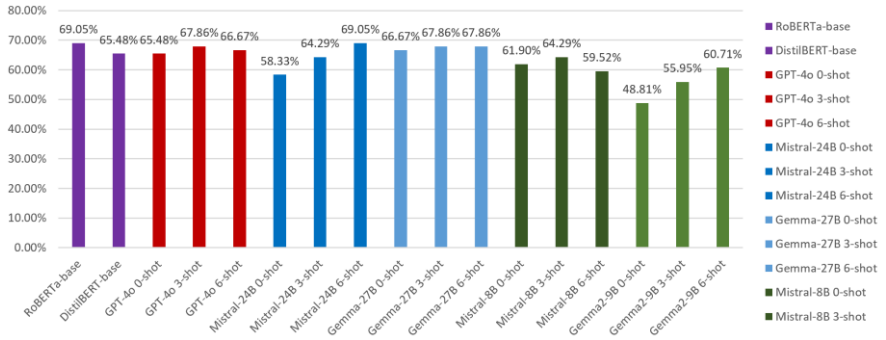

Fig. 4. Micro-F1 scores of encoder-based and decoder-based models on the ECC dataset.

Table 5 and Figure 2 summarize the performance of decoder-based models on the ECC dataset under 0-shot, 3-shot, and 6-shot prompt settings. Overall, most models benefited from few-shot prompting, especially from 0-shot to 3-shot. However, performance gains beyond 3-shot were not always consistent.

Among all models, Mistral-24B achieved the highest Micro-F1 (69.05%) and Macro-F1 (64.43%) under the 6-shot setting, even outperforming GPT-4o. This demonstrates the strong potential of medium-sized decoder models when appropriately prompted. Gemma-27B also showed stable and competitive results across all prompt configurations, consistently maintaining Macro-F1 scores above 60%.

In contrast, small models such as Mistral-8B and Gemma2-9B performed worse, especially in the 6-shot setting, with noticeable declines in Macro-F1 scores. This suggests that smaller models may struggle with complex prompts or have limited capacity for temporal reasoning.

These findings highlight the importance of model size and prompt design. While large models like GPT-4o are strong performers, medium-sized models such as Mistral-24B can achieve comparable or even better results when given well-designed prompts. This suggests that Mistral-24B offers a promising combination of efficiency and accuracy, making it a competitive alternative to larger models in certain scenarios.

Table 6. Prompt-Based Decoder Model Performance on the Social Media Dataset (Chinese).

Model	Size	Prompt Setting	Micro-F1	Macro-F1
GPT-4o	Large	0-shot	38.70%	33.95%
GPT-4o	Large	3-shot	47.32%	44.76%
GPT-4o	Large	6-shot	**51.43%**	48.48%
Mistral-24B	Medium	0-shot	14.38%	11.12%
Mistral-24B	Medium	3-shot	28.14%	28.90%
Mistral-24B	Medium	6-shot	34.19%	34.61%
Gemma-27B	Medium	0-shot	23.06%	18.54%
Gemma-27B	Medium	3-shot	38.87%	34.39%
Gemma-27B	Medium	6-shot	**47.66%**	39.48%
Mistral-8B	Small	0-shot	9.00%	6.00%
Mistral-8B	Small	3-shot	18.00%	21.00%
Mistral-8B	Small	6-shot	23.97%	25.93%
Gemma2-9B	Small	0-shot	20.26%	22.10%
Gemma2-9B	Small	3-shot	**40.35%**	38.25%
Gemma2-9B	Small	6-shot	40.24%	37.31%

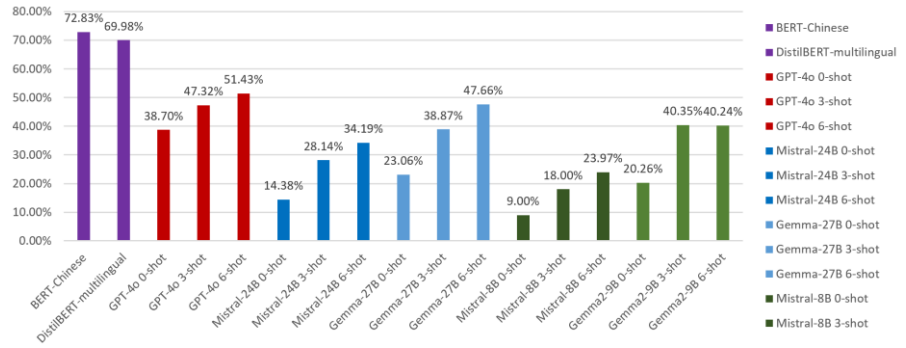

Fig. 5. Micro-F1 scores of encoder- and decoder-based models on the Social Media Dataset.

Table 6 and Figure 3 summarize decoder-based model performance on the Social Media dataset across different prompt settings. Most models improved with more in-context examples, confirming the effectiveness of few-shot learning in financial temporal reasoning.

GPT-4o, the largest model, achieved the best results with a Micro-F1 of 51.43% and Macro-F1 of 48.48% under the 6-shot setting. Among medium-sized models, Gemma-27B showed notable improvement from 23.06% to 47.66% in Micro-F1, narrowing the gap with GPT-4o. Mistral-24B also improved but consistently performed below Gemma-27B.

Small models such as Mistral-8B and Gemma2-9B had lower overall scores. Mistral-8B began with 9.00% Micro-F1 and reached 23.97% in the 6-shot setting. In

comparison, Gemma2-9B exceeded 40% Micro-F1 at both 3-shot and 6-shot, suggesting smaller models can still benefit from well-crafted prompts.

These results suggest that while large models like GPT-4o lead in performance, medium-sized models such as Gemma-27B can offer comparable results with appropriate prompting and greater efficiency.

4.3 Best Performing Settings Summary

Table 7. Top Performing Models on the ECC Dataset.

Rank	Model Type	Model	Prompt Setting	Micro-F1	Macro-F1
1	Encoder	RoBERTa-base	Fine-tuned	69.05%	67.06%
2	Decoder (Medium)	Mistral-24B	6-shot	69.05%	64.43%
3	Decoder (Medium)	Gemma-27B	3-shot	67.86%	64.50%
4	Decoder (Large)	GPT-4o	3-shot	67.86%	62.36%
5	Decoder (Medium)	Gemma-27B	6-shot	67.86%	61.36%

Table 8. Top Performing Models on the Social Media Dataset.

Rank	Model Type	Model	Prompt Setting	Micro-F1	Macro-F1
1	Encoder	BERT-Chinese	Fine-tuned	72.83%	53.40%
2	Encoder	DistilBERT-multilingual	Fine-tuned	69.98%	53.50%
3	Decoder (Large)	GPT-4o	6-shot	51.43%	48.48%
4	Decoder (Medium)	Gemma-27B	6-shot	47.66%	39.48%
5	Decoder (Large)	GPT-4o	3-shot	47.32%	44.76%

Tables 7 and 8 present the top-performing models across both datasets. On the ECC dataset, the fine-tuned RoBERTa-base and the 6-shot Mistral-24B decoder tied with the highest Micro-F1 of 69.05%, showing that well-prompted decoder models can match encoder-based models in English tasks.

On the Social Media dataset, fine-tuned encoders like BERT-Chinese and DistilBERT-multilingual clearly outperformed decoder-based models. This performance gap may reflect the limited Chinese-language capabilities of decoder models not trained on domain-specific data.

These findings suggest that while decoder models can compete with encoders in English under effective prompting, encoder-based models remain superior for Chinese-language financial classification tasks.

4.4 Analysis and Discussion

Class Imbalance and Metric Gap. Figure 4 compares the Micro and Macro F1 gaps across models and datasets. On the ECC dataset, encoder-based models exhibit rela-

tively smaller gaps, while decoder-based models show more variation. This may be due to class balancing applied during ECC preprocessing, which helps encoders better capture minority classes. On the Social Media dataset, where no balancing was applied, the encoder models show large gaps, indicating difficulties in handling unbalanced labels. In contrast, decoder models tend to maintain smaller gaps, even without additional balancing. This suggests that prompt-based models may inherently handle label imbalance better, possibly due to the contextual information provided through in-context examples.

These findings highlight how data characteristics and modeling strategies jointly influence performance in imbalanced classification tasks.

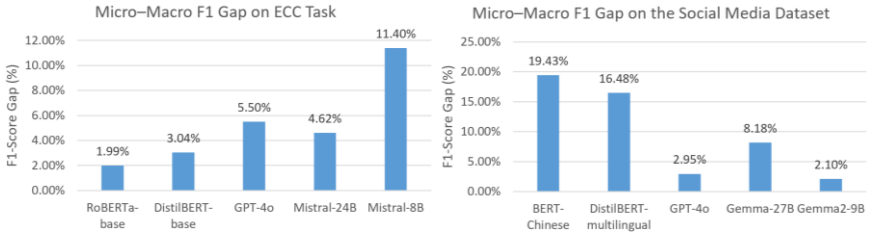

Fig. 6. Micro–Macro F1 Gap on Social Media Task on the ECC and Social Media Dataset.

Error Analysis. Preliminary error analysis suggests that both encoder- and decoder-based models struggle with ambiguous or underspecified temporal expressions, particularly in minority classes. While decoder models benefit from few-shot prompts, vague or overly generic examples may lead to incorrect generalizations. This highlights the potential benefit of incorporating expert-verified, task-specific instructions into prompts, rather than relying solely on example-based demonstrations.

RQ1: Comparison of Fine-Tuned Encoder Models and Prompt-Based Decoder Models. Across both datasets, fine-tuned encoder-based models demonstrated strong and stable performance, particularly in Chinese-language tasks. Decoder-based models, when paired with well-designed prompts, were able to match or even exceed encoder performance in English tasks. These results indicate that decoder models are viable alternatives in few-shot scenarios, especially when fine-tuning resources are limited or when the target language is well-supported..

RQ2: Performance of Small vs. Large Language Models under Prompt-Based Settings. Model size had a significant impact on decoder performance. Larger and medium-sized models consistently outperformed smaller ones across most prompt settings. However, some medium-sized models, such as Mistral-24B and Gemma-27B, achieved results comparable to large models like GPT-4o under effective prompting. This suggests that with optimized prompt design, medium-sized models can offer a strong balance between computational efficiency and predictive accuracy..

RQ3: Model Behaviors in English and Chinese Tasks. Language characteristics played a crucial role in model behavior. In English tasks, decoder models showed competitive performance, benefiting from prompt-based learning. In contrast, their performance on Chinese tasks lagged behind that of fine-tuned encoders. This disparity may be attributed to differences in pretraining data coverage, language-specific representations, or the effectiveness of prompts in each language. These findings underscore the need for language-aware model selection and prompt design in multilingual classification tasks.

5 Conclusions

This study compared fine-tuned encoder models and prompt-based decoder models for temporal reasoning in financial texts across English and Chinese datasets. Fine-tuned encoders such as RoBERTa-base and BERT-Chinese consistently delivered strong performance when labeled data was available. Meanwhile, medium-sized decoder models like Mistral-24B and Gemma-27B achieved competitive results under few-shot prompting, offering a promising alternative when fine-tuning is not feasible. Although small models performed poorly overall, some medium-sized models approached or matched larger ones, particularly on the English dataset.

This work contributes to the understanding of how model architecture, scale, and training strategy affect performance in financial temporal reasoning tasks. It provides comparative insights on fine-tuning versus prompting and extends evaluation to multilingual and imbalanced datasets. One key observation is that decoder models demonstrate greater resilience in handling minority classes under imbalanced conditions, even without data balancing techniques.

For practitioners, the results suggest that fine-tuned encoder models remain a reliable choice when annotated data is available. However, prompt-based models, especially medium-sized ones, provide an effective balance between performance and resource demands. These findings offer practical guidance for model selection in real-world deployments, particularly in resource-constrained environments where multilingual understanding and label imbalance are common.

Future work may explore more advanced prompting techniques such as instruction tuning, investigate combined fine-tuning and prompting strategies, and expand analysis to other languages and domains. A deeper investigation into the architectural factors and pretraining data alignment that influences model behavior may also shed light on performance variations across languages and tasks. In addition, incorporating task-specific guidance or domain expertise into prompt design may help improve few-shot model generalization, particularly in cases involving complex or ambiguous temporal expressions.

6 Acknowledgement

This research was supported by the Industrial Technology Research Institute (ITRI) and National Taipei University (NTPU), Taiwan, under grants NTPU-114A513E01

and NTPU-113A513E01; the National Science and Technology Council (NSTC), Taiwan, under grant NSTC 113-2425-H-305-003; the ATEC Group under grant NTPU-112A413E01; and National Taipei University (NTPU) under grant 114-NTPU_ORDA-F-004.

References

1. Alhamzeh A. Financial argument quality assessment in earnings conference calls. International Conference on Database and Expert Systems Applications: Springer; 2023. p. 65-81.
2. Chiu Jr C, Chen C-C, Huang H-H, Chen H-H. Pre-Finetuning with Impact Duration Awareness for Stock Movement Prediction. arXiv preprint arXiv:240917419. 2024.
3. Zong S, Ritter A, Hovy E. Measuring forecasting skill from text. arXiv preprint arXiv:200607425. 2020.
4. Kenton JDM-WC, Toutanova LK. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of naacL-HLT: Minneapolis, Minnesota; 2019. p. 2.
5. Sun C, Qiu X, Xu Y, Huang X. How to fine-tune bert for text classification? Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18: Springer; 2019. p. 194-206.
6. Lin C-Y, Chen C-C, Huang H-H, Chen H-H. Argument-Based Sentiment Analysis on Forward-Looking Statements. Findings of the Association for Computational Linguistics ACL 20242024. p. 13804-15.
7. Pan L, Hang C-W, Sil A, Potdar S. Improved text classification via contrastive adversarial training. Proceedings of the AAAI Conference on Artificial Intelligence2022. p. 11130-8.
8. Jiang T, Duan J, Li W, Zhang M, Liu Y, Yang J. A Study of Risk Prediction Based on a Hybrid Model of LSTM and BERT. 2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE): IEEE; 2024. p. 1321-4.
9. Mayer CW, Ludwig S, Brandt S. Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models. Journal of Research on Technology in Education. 2023;55(1):125-41.
10. Loukas L, Stogiannidis I, Malakasiotis P, Vassos S. Breaking the bank with ChatGPT: few-shot text classification for finance. arXiv preprint arXiv:230814634. 2023.
11. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:230211382. 2023.
12. Li Y, Wang S, Ding H, Chen H. Large language models in finance: A survey. Proceedings of the fourth ACM international conference on AI in finance2023. p. 374-82.
13. Zhang Q, Liu Z, Pan S. The rise of small language models. IEEE Intelligent Systems. 2025;40(1):30-7.
14. Recasens PG, Zhu Y, Wang C, Lee EK, Tardieu O, Youssef A, et al. Towards Pareto optimal throughput in small language model serving. Proceedings of the 4th Workshop on Machine Learning and Systems2024. p. 144-52.
15. Yousri R, Safwat S. How Big Can It Get? A comparative analysis of LLMs in architecture and scaling. 2023 International Conference on Computer and Applications (ICCA): IEEE; 2023. p. 1-5.