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





Bor-Jen Chen  Wen-Hsin Hsiao  Hsin-Ting Lu  Min-Yuh Day*

Graduate Institute of Information Management, National Taipei University, New Taipei City, Taiwan
myday@gm.ntpu.edu.tw*



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



The 16th International Workshop on Mining and Analyzing Social Networks for Decision Support (MSNDS 2025), in Proceedings of the 2025 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2025), Niagara Falls, Ontario, Canada, August 25-28, 2025.

# Outline

1. Introduction
2. Related Work
3. Methodology
4. Experimental Results and Analysis
5. Conclusions

# 1. Introduction

- Research Background
- Research Motivation
- Research Objectives

# Research Background

- Rapid progress in **NLP and LLMs** opens new opportunities in financial text analysis.
- Financial texts (earnings conference calls, social media) often contain **time-sensitive insights** crucial for market evaluation.
- **Challenge:** Temporal reasoning is hard due to vague expressions (e.g., “soon”, “in the near future”), varied patterns, and multilingual complexities.
- Need for models that can extract **accurate temporal information** with strong contextual understanding.

# Research Motivation

- Current models have difficulty handling **implicit temporal expressions** in finance.
- Two main approaches:
  - **Encoder-based fine-tuning** (e.g., RoBERTa, BERT) → strong with labeled data.
  - **Decoder-based prompting** (e.g., GPT-4o, Mistral, Gemma) → flexible in few-shot/zero-shot.

**Research Gap**

Lack of systematic comparison across languages, model sizes, and prompt strategies.

# Research Objectives

## Main Objectives

Compare **encoder-based fine-tuning** and **decoder-based prompting**, and try different **model sizes**.

Use two datasets:

- **Earnings Conference Call (ECC) – English:** formal transcripts from top tech companies' quarterly calls, containing managerial forecasts and analyst Q&A.
- **Social media – Chinese:** investor opinions from an online forum.

## Research Questions

RQ1: How do **fine-tuned encoder** compare with **prompt-based decoder** models in temporal classification?

RQ2: Can **small/mid-sized models** compete with large models?

RQ3: How does **language difference** affect performance?

# 2. Related Work

- Temporal Reasoning in Finance
- Fine-Tuning Strategies for Encoder Models
- Prompting Methods with Decoder Models
- Model Scaling and Small Language Models

# Temporal Reasoning in Finance

| Contribution | Limitation / Challenge | Reference |
|---|---|---|
| Built ECC dataset with temporal references; showed temporal framing affects persuasiveness. | Focused only on English transcripts. | Alhamzeh (2023) |
| Introduced impact duration awareness for stock prediction; modeled how long news remains relevant. | Based on financial news; less applicable to informal texts. | Chiu et al. (2024) |
| Linked forward-looking text to real outcomes; measured forecasting skill. | Variation in predictive value, hard to generalize. | Zong et al. (2020) |

*These studies highlight that financial temporal reasoning requires not only detecting expressions but also modeling duration and predictive validity.*

Source: This study

# Fine-Tuning Strategies for Encoder Models

| Contribution | Limitation / Challenge | Reference |
|---|---|---|
| BERT as a pre-trained encoder, strong baseline for NLP. | Needs large labeled data for downstream tasks. | Devlin et al. (2019) |
| Showed hyperparameter tuning (batch, lr, epochs) critical for BERT fine-tuning. | Sensitive to parameter choices. | Sun et al. (2019) |
| Fine-tuned FinBERT for argument-based sentiment in financial texts. | Domain-specific; limited cross-task generalization. | Lin et al. (2024) |
| Contrastive adversarial training improves robustness and generalization. | Requires extra data processing. | Pan et al. (2022) |
| Combined BERT with LSTM for financial risk prediction. | Higher complexity, harder deployment. | Jiang et al. (2024) |

*Fine-tuning remains powerful but requires labeled data and careful adaptation to domain-specific tasks.*

Source: This study

# Prompting Methods with Decoder Models

| Contribution | Limitation / Challenge | Reference |
|---|---|---|
| Provided overview of prompt-based learning (zero-shot, few-shot, CoT). | Conceptual, limited financial application. | Mayer et al. (2023) |
| Tested ChatGPT with few-shot prompts for finance (sentiment, stance, topic). Competitive without fine-tuning. | Performance varies with prompt design. | Loukas et al. (2023) |
| Introduced prompt pattern catalog (reasoning, role prompting, examples). | Requires expert knowledge for effective design. | White et al. (2023) |

*Prompting offers flexibility and strong performance, but effectiveness depends heavily on prompt design.*

Source: This study

# Model Scaling and Small Language Models

| Contribution | Limitation / Challenge | Reference |
|---|---|---|
| Surveyed LLMs in finance; emphasized resource demands. | Large-scale models costly to deploy. | Li et al. (2023) |
| Highlighted rise of Small Language Models (SLMs), optimized with quantization and instruction tuning. | May lag behind LLMs in very complex tasks. | Zhang et al. (2025) |
| Proposed serving strategies for SLMs, achieving Pareto-optimal throughput with minimal accuracy loss. | Focused on system-level optimization, not NLP tasks. | Recasens et al. (2024) |
| Analyzed scaling laws; model size alone does not guarantee better performance. | Scaling insights mostly outside finance. | Yousri & Safwat (2023) |

*Smaller models with optimization show cost-effective alternatives, challenging the assumption that bigger is always better.*

Source: This study

# Summary & Research Gap

**Summary**

Prior studies explored temporal reasoning, encoder fine-tuning, and decoder prompting. LLM scaling shows strong performance, while smaller models are emerging as efficient alternatives.

**Research Gap**

- Limited systematic **comparison between encoder fine-tuning and decoder prompting** in financial temporal reasoning.
- Limited exploration of **mid-sized models** as cost-effective options.
- Lack of **multilingual evaluation** (most work focused on English).
- Insufficient insights into **prompt design strategies** for financial contexts.

*This study aims to fill these gaps by comparing encoder and decoder approaches across model sizes and languages, using both English (ECC) and Chinese (Social Media) datasets.*

Source: This study

# 3. Methodology

- Research Framework
- Datasets and Preprocessing
- Model Selections
- Training and Inference Procedure
- Evaluation Metrics

# Research Framework

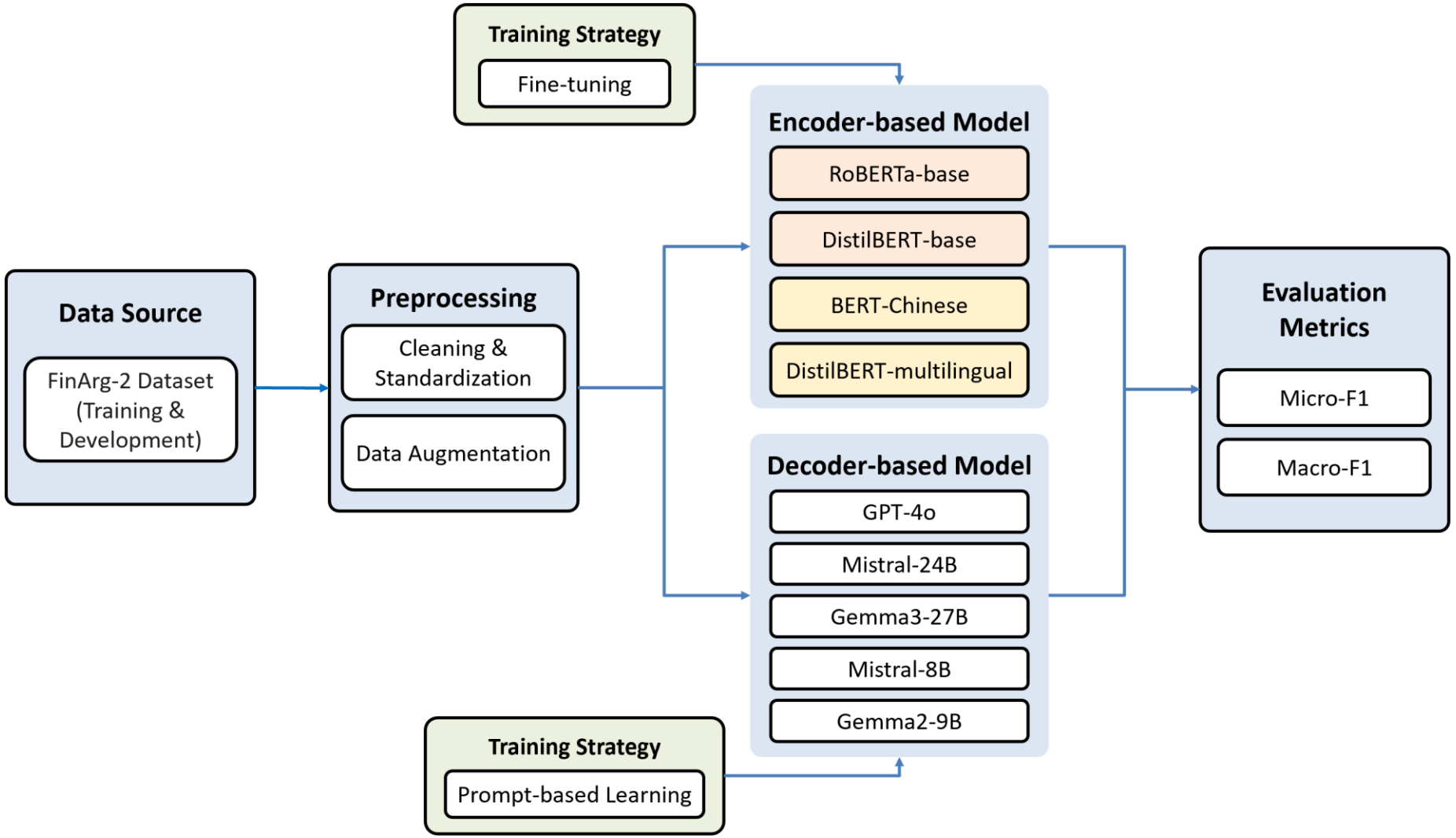



Fig. 1. Overall research framework for temporal reasoning with fine-tuned encoders and prompt-based decoders.

Source: This study

# Research Framework

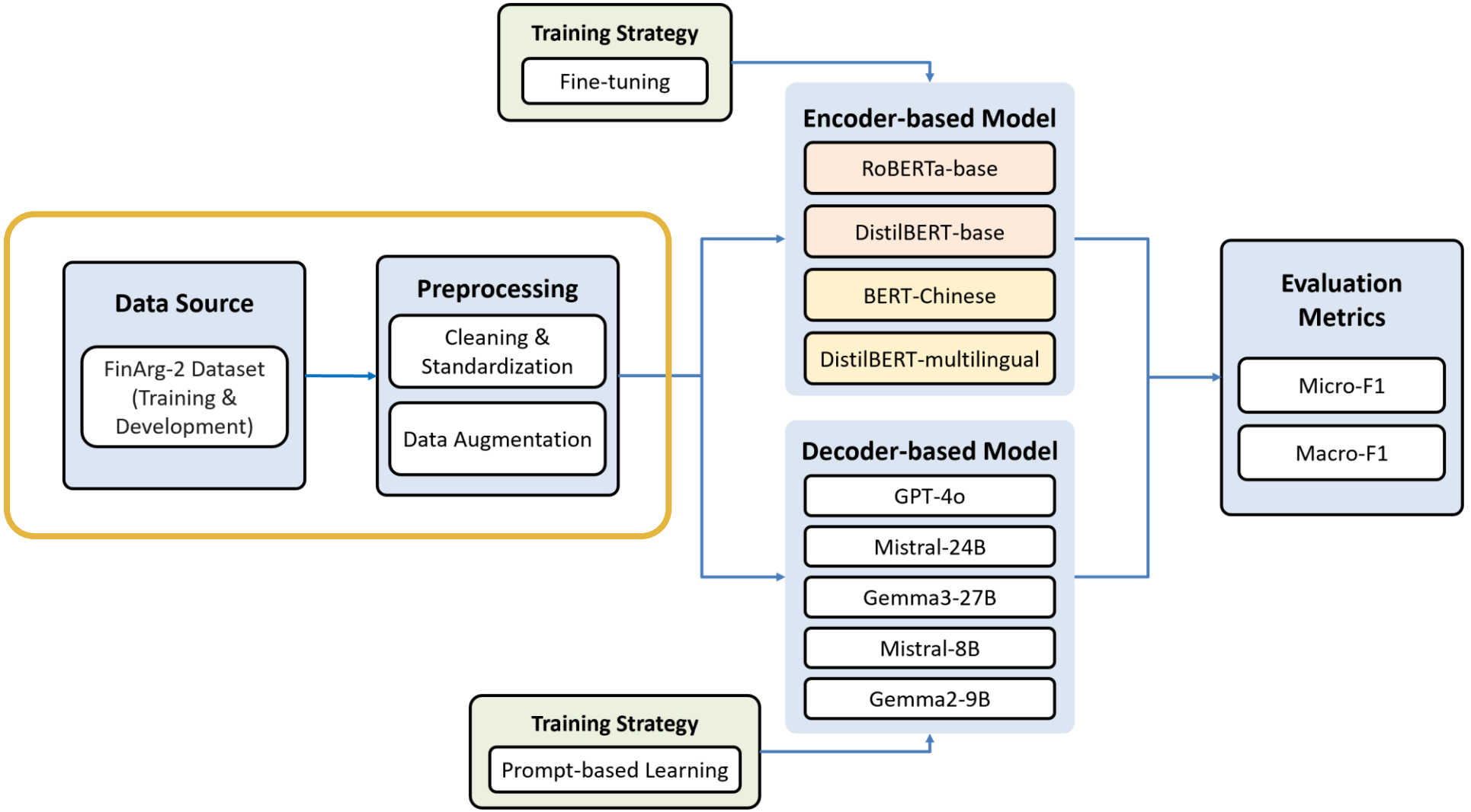



Fig. 1. Overall research framework for temporal reasoning with fine-tuned encoders and prompt-based decoders.

Source: This study

# ECC Dataset (English)

**Source**: Financial Modeling Prep API (2015–2019).
**Coverage**: 80 transcripts from top tech firms (Amazon, Apple, Microsoft, Facebook).
**Labels**:

- 0: No time reference
- 1: Long past (> 6 months)
- 2: Short past (≤ 6 months, e.g., this or next quarter)

**Dataset example:**
**claim_text**：claim: So certainly, a lot going on in international, a lot that's really good, adding Prime subscribers at a high clip, continuing to add selection at FBA sellers.
**premise_texts**：""So you'll see devices, you see video content."", 'It may be getting there a little slower than the starting point in the U.S. but we see it really showing up in customer engagement and customer purchases.', 'On the AWS side, I think the 2016 to 2015 comparison probably stands on its own and 2014 falls by the wayside, so I would encourage you to look at recent trends.', ""So it's the whole array of Prime offering, Prime Now, Same-Day.""]
**year:** 2016
**quarter:** Q1
**label:** 1

# Tasks - ECC (Earnings Conference Call) Dataset

Managers or analysts often mention **past events** in transcripts.

“Sales were strong last quarter” → **Short past** (within 6 months)

“Last year we faced supply issues” → **Long past** (over 6 months)

Some statements have **no time reference**

→ Task: classify into **short past**, **long past**, or **no time reference**

# Social Media Dataset (Chinese)

**Source**: Mobile01 investment forum (8,760 posts).
**Nature**: Informal investor opinions about stock trends & news.
**Annotation**: By financial experts; reliability checked with Cohen's Kappa (≈ 66%).
**Labels**:

- Within 1 week
- Longer than 1 week
- Unsure

**Dataset example :**
**'text':** '看來中華電還是比定存好\n \n 等低點來買一些好了',
(It seems that Chunghwa Telecom is still better than fixed deposits. \n \n I'd better wait for the price to drop and buy some.)
**'Label_duration':** 'Longer than 1 week',

# Tasks - Social Media Dataset (Investor Forum)

Investor posts often imply **how soon** something may happen.

“The price may rise tomorrow” → **Within one week**

“This stock has long-term growth” → **Longer than one week**

“Maybe soon…” → **Unsure**

→ Task: classify into **within one week**, **longer than one week**, or **unsure**

# ECC Training Dataset

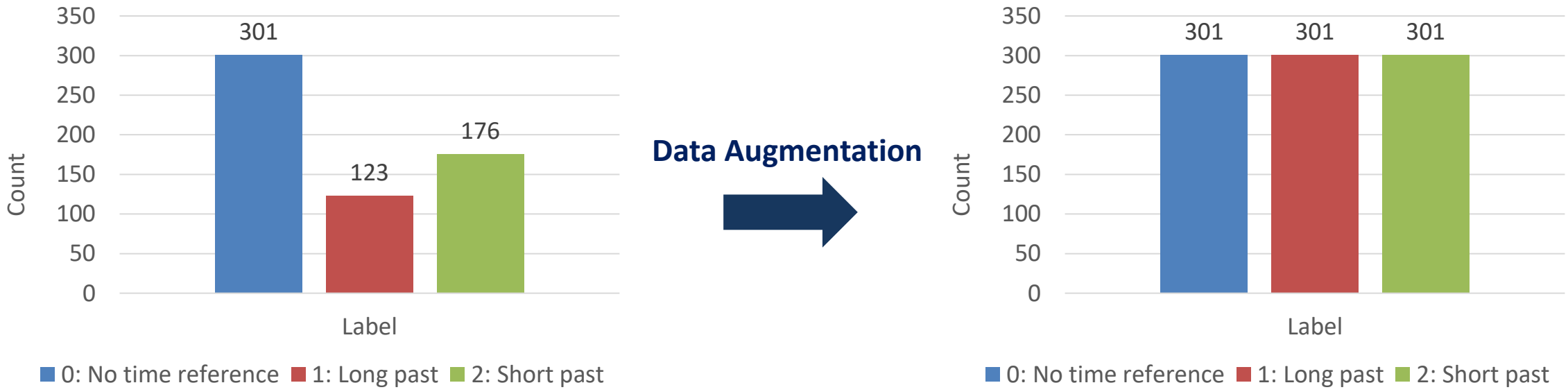
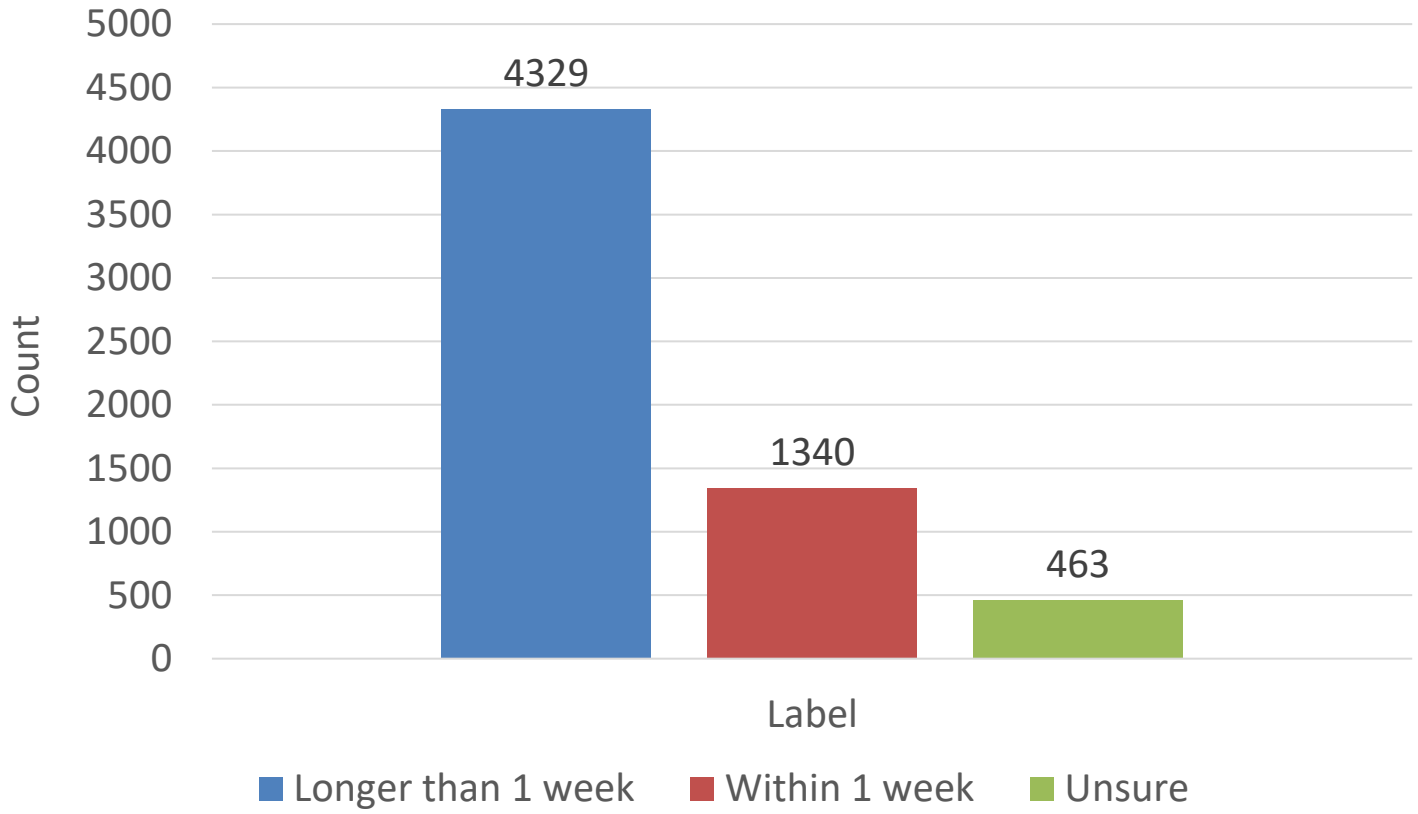



Fig. 2. Label distribution diagram before and after data argumentation on ECC Training Dataset.

Source: This study

# Social Media Training Dataset




**No Data Augmentation**

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

Source: This study

# Research Framework

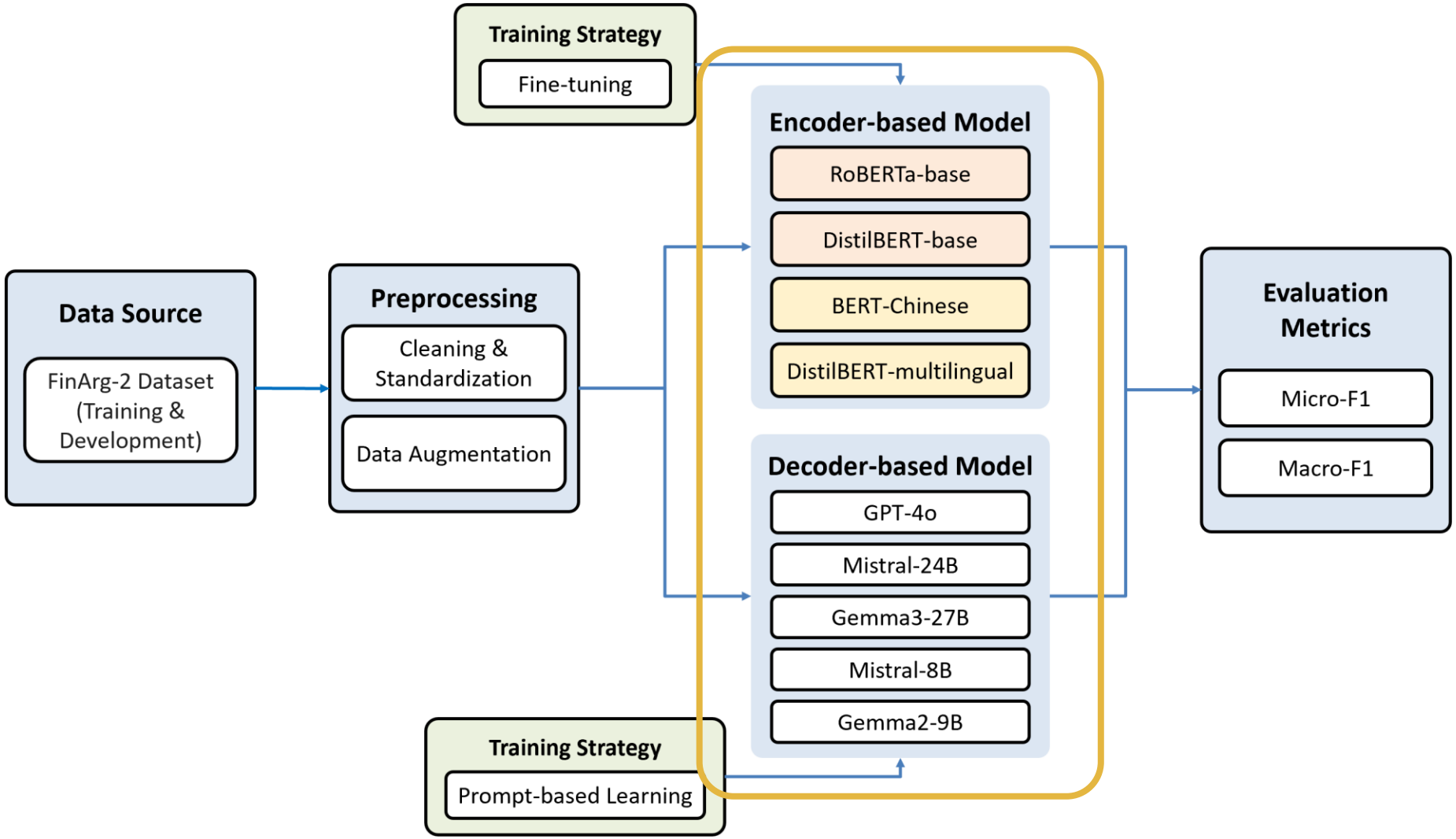



Fig. 1. Overall research framework for temporal reasoning with fine-tuned encoders and prompt-based decoders.

Source: This study

# Model Selections - Encoder Models

| 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. |

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

Source: This study

# Model Selections – Decoder Models

| 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 |

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

Source: This study

# Research Framework

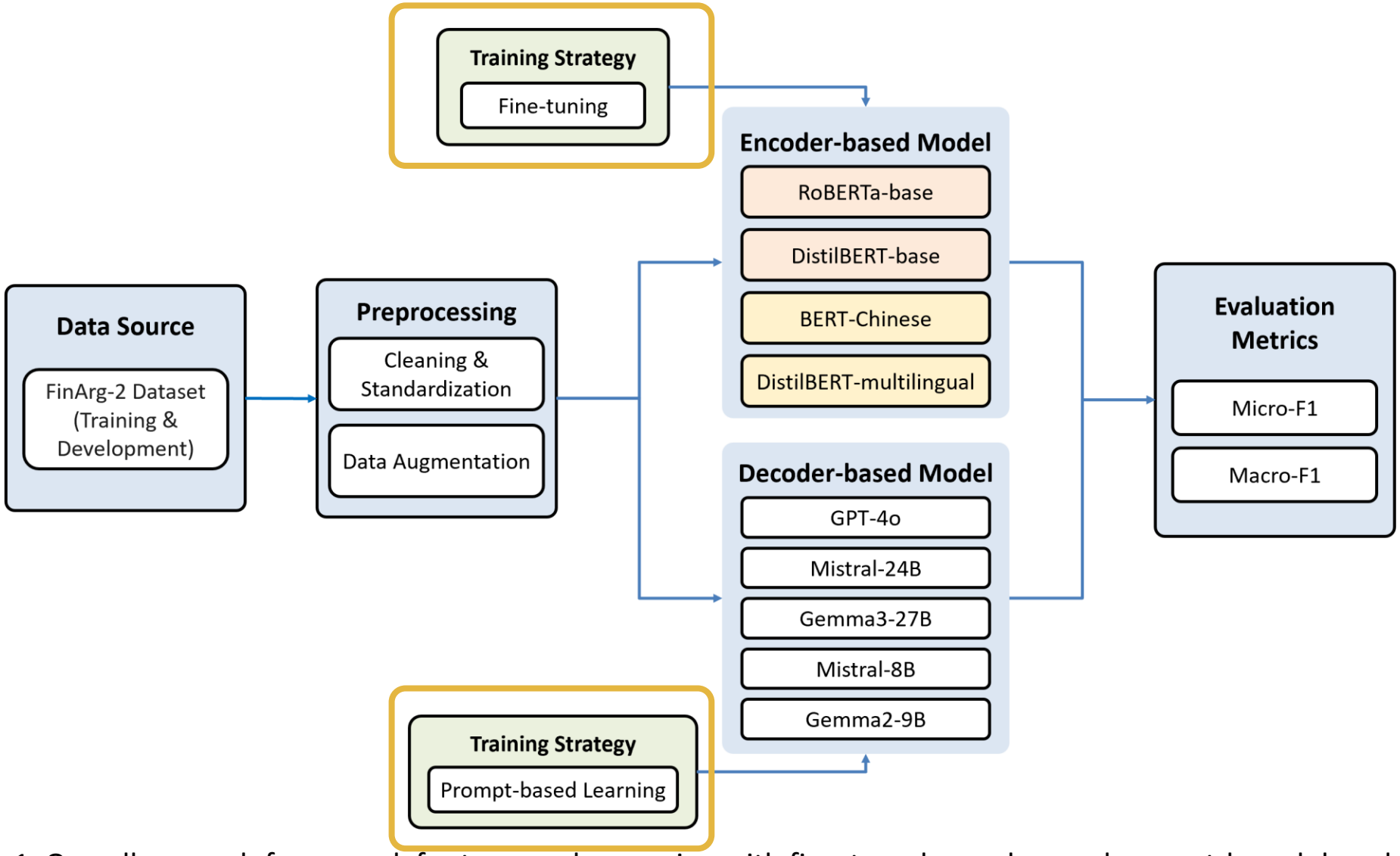



Fig. 1. Overall research framework for temporal reasoning with fine-tuned encoders and prompt-based decoders.

Source: This study

# Encoder Fine-Tuning

**Setup**

- Transformer-based encoders trained with supervised classification
- Tokenization: [CLS], [SEP], truncation/padding (128 or 256 tokens).
- Optimization: AdamW + Cross-Entropy Loss.
- Training: 3–6 epochs, early stopping on Micro/Macro-F1.
- Regularization: gradient clipping, weight decay (0.01), dropout.

| 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 |

Table 3. Fine-Tuning Hyperparameter Settings.

Source: This study

# Decoder Inference Settings

**Prompt-Only Inference**

No parameter updates; evaluated under in-context learning.

**Shot settings**:

1. Zero-shot
2. 3-shot → one example per class
3. 6-shot → two examples per class

**Prompt Structure**

1. Task instruction (prediction goal)
2. Label definitions (class mappings)
3. In-context examples (from correctly predicted training data)
4. Test input ending with *"Label:"*

# Research Framework

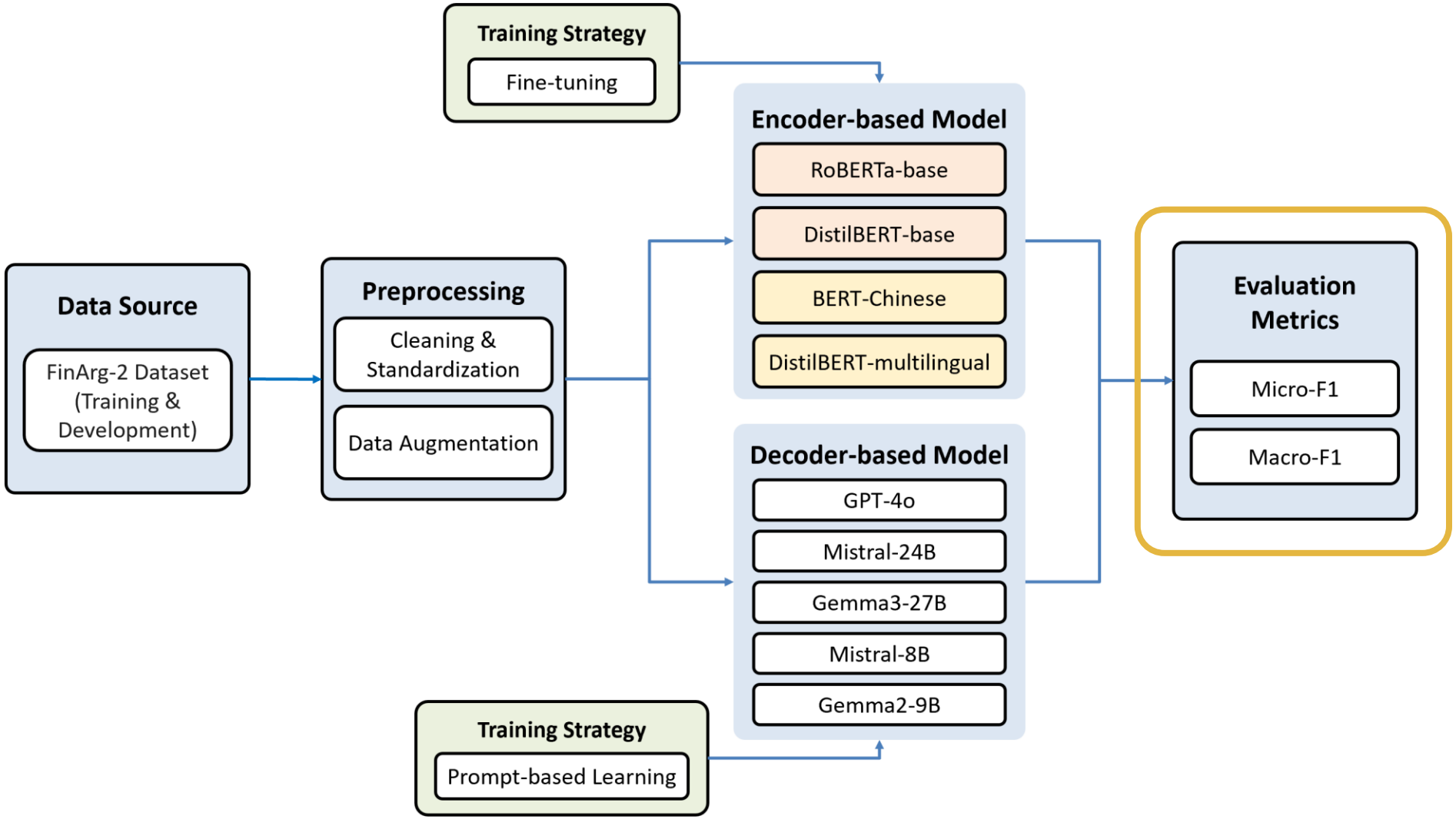



Fig. 1. Overall research framework for temporal reasoning with fine-tuned encoders and prompt-based decoders.

Source: This study

# Evaluation Metrics

**Micro-F1**

- Aggregates across all predictions
- Captures overall accuracy, but favors majority classes

**Macro-F1**

- Averages per-class F1-scores
- Treats all classes equally, giving fair weight to minority classes.

→ Provide a balanced view:
Overall predictive ability and robustness under class imbalance

# 4. Data Analysis and Discussion

- Encoder-Based Model Performance
- Decoder-Based Model Performance
- Best Performing Settings Summary
- Analysis and Discussion

# Encoder-Based Model 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. Encoder-Based Models Performance.

Source: This study

# Decoder-Based Model on ECC Dataset (English)

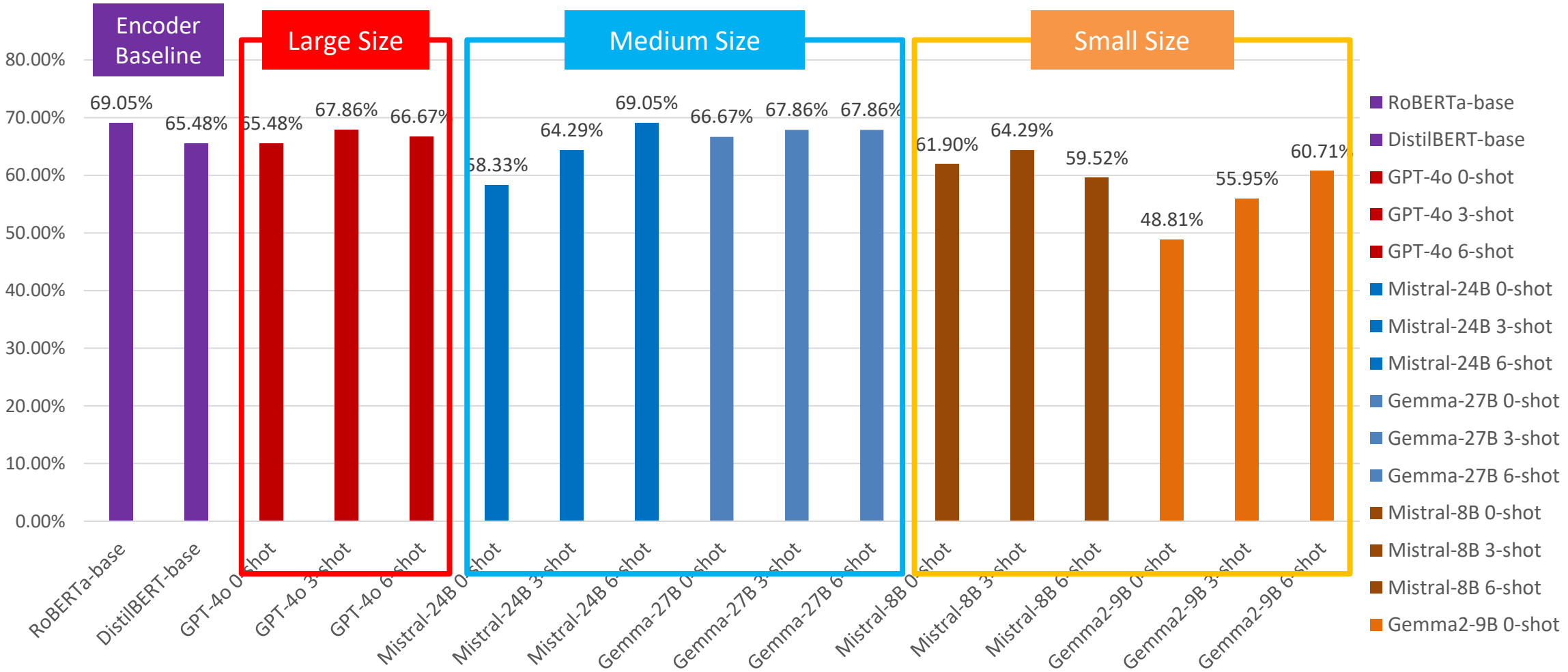



Fig. 2. Micro-F1 scores of encoder-based and decoder-based models on the ECC dataset.

Source: This study

# Decoder-Based Model on ECC Dataset (English)

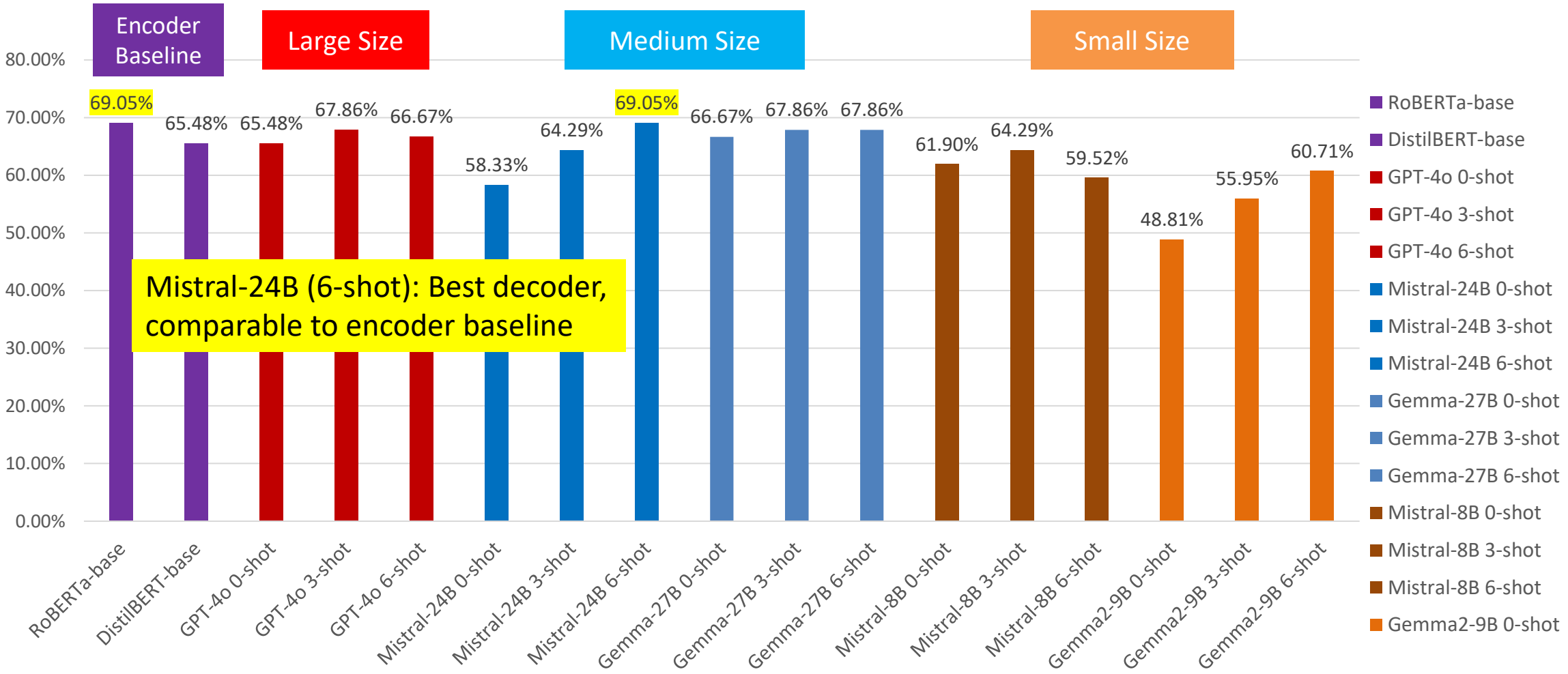



Fig. 2. Micro-F1 scores of encoder-based and decoder-based models on the ECC dataset.

Source: This study

# Best Performing Settings Summary (ECC)

| 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 7. Top Performing Models on the ECC Dataset.

Source: This study

# Decoder-Based Model on Social Media Dataset (Chinese)

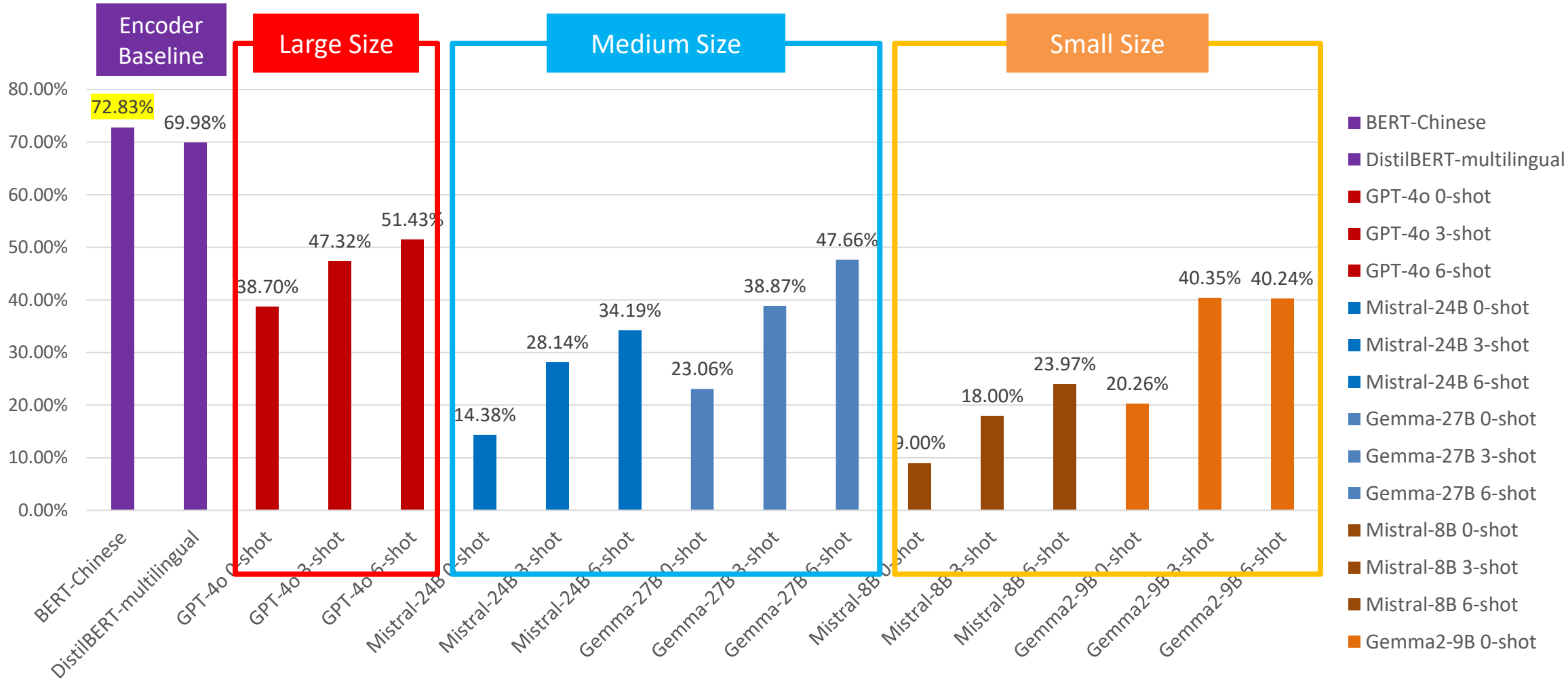



Fig. 3. Micro-F1 scores of encoder- and decoder-based models on the Social Media Dataset.

Source: This study

# Decoder-Based Model on Social Media Dataset (Chinese)

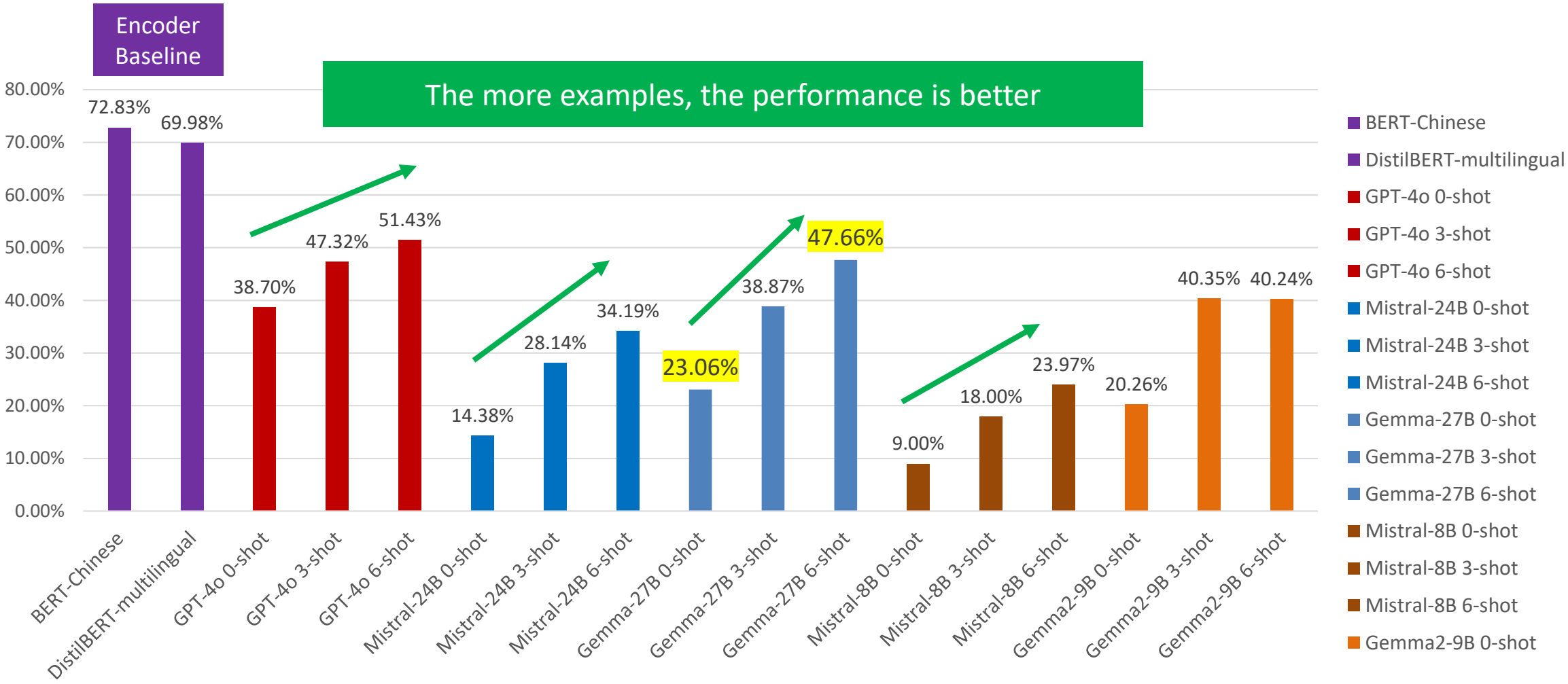



Fig. 3. Micro-F1 scores of encoder- and decoder-based models on the Social Media Dataset.

Source: This study

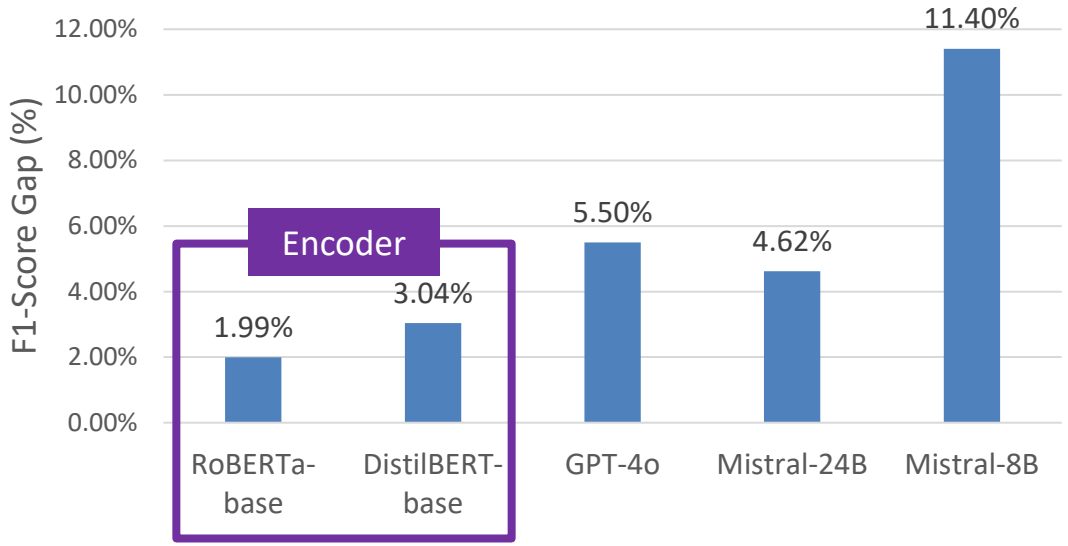
# Best Performing Settings Summary (Social Media)

| 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% |

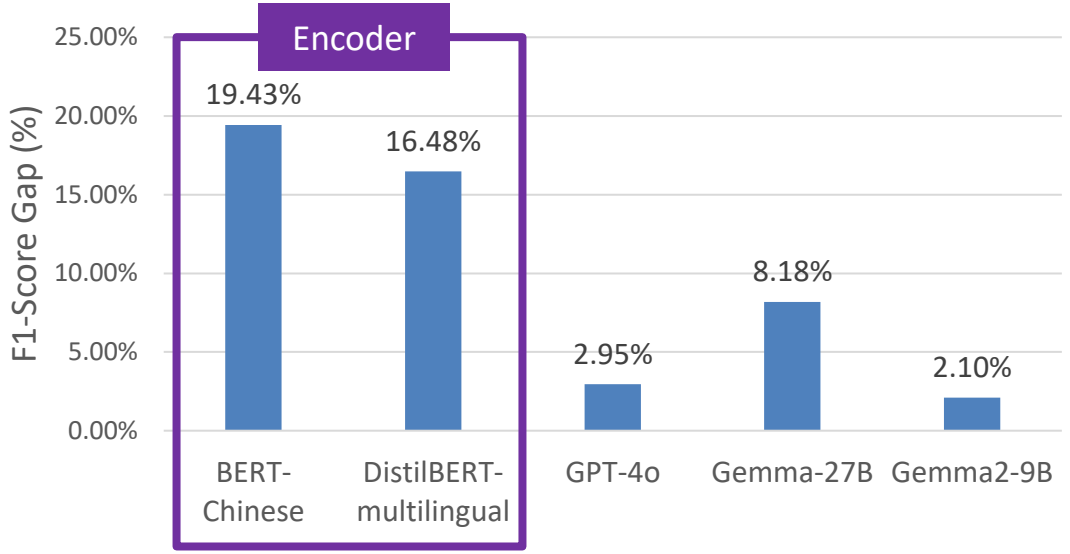
Table 8. Top Performing Models on the Social Media Dataset.

Source: This study

# Class Imbalance and Metric Gap




On ECC (**with balancing**)
Encoder models show smaller Micro–Macro F1 gaps.




On Social Media (**no balancing**)
Encoder gaps widen, while decoder models remain relatively stable.

→ Prompt-based models may inherently handle label imbalance better

Fig. 4. Micro–Macro F1 Gap on Social Media Task on the ECC and Social Media Dataset.

Source: This study

# RQ1: Comparison of Fine-Tuned Encoder Models and Prompt-Based Decoder Models.

- Fine-tuned encoders: strong & stable performance, especially in Chinese tasks.
- Decoders with well-designed prompts: competitive in ECC dataset, sometimes exceeding encoders.

→ Decoders are viable alternatives in few-shot or resource-limited scenarios.

# RQ2: Performance of Small vs. Large Language Models under Prompt-Based Settings.

- Larger & medium-sized models consistently outperform smaller ones.
- Medium models (Mistral-24B, Gemma-27B) can match large models (GPT-4o) with good prompt.

→ Medium models offer a balance of efficiency and accuracy.

# RQ3: Model Behaviors in English and Chinese Tasks.

- In English: decoders with prompt-based learning perform close to encoders.
- In Chinese: encoders clearly outperform decoders.
- Cause: limited Chinese pretraining resources and domain-specific gaps.

→ Need for **language-aware prompt design** and specialized training for Chinese financial texts.

# 5. Conclusions

- Research Contributions
- Managerial Implications
- Future work

# Research Contributions

- Comparative study of model architecture, scale, and training strategy on financial temporal reasoning.
- Evaluation across **English & Chinese datasets**, including balanced and imbalanced datasets.
- **Decoder models show resilience on minority classes**, even without balancing.

# Managerial Implications

- **Encoders:** reliable when annotated data is available → best choice for **high-accuracy tasks**.
- **Medium-sized decoders: efficient** and good for **fast deployment** in **resource-limited settings** or when labeled data is scarce.
- Provides **practical guidance**: organizations can choose encoders for stability or **decoders for scalability and cost-effectiveness**.

# Future work

- Explore **advanced prompting** (e.g., instruction tuning).
- Incorporate **domain expertise** into prompt design for better generalization.
- Investigate hybrid strategies or **muti-agent** collaboration.
- Extend to **other languages and financial domains**.

# 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.

Thanks for Listening!

Q&A

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




Bor-Jen Chen  Wen-Hsin Hsiao  Hsin-Ting Lu  Min-Yuh Day*

Graduate Institute of Information Management, National Taipei University, New Taipei City, Taiwan
myday@gm.ntpu.edu.tw*

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

The 16th International Workshop on Mining and Analyzing Social Networks for Decision Support (MSNDS 2025), in Proceedings of the 2025 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2025), Niagara Falls, Ontario, Canada, August 25-28, 2025.