EFSA-CLC: Enhancing Zero-shot Entity-level Financial Sentiment Analysis with Cross-lingual Collaboration

Senbin Zhu[1], Hongde Liu[1], Chenyuan He[1], Zhangzu Geng[2], Yuxiang Jia[1]*
[1]School of Computer and Artificial Intelligence, Zhengzhou University, China
[2]Guangdong Experimental High School, China
{nlpbin,lhd_1013,hechenyuan_nlp}@gs.zzu.edu.cn, gengzhangzu@qq.com
Correspondence: ieyxjia@zzu.edu.cn

Abstract

Entity-level sentiment analysis is becoming increasingly important in the context of diverse financial texts, and large language models demonstrate significant potential under zero-shot settings. While it is well recognized that different languages embody distinct cognitive patterns, the use of multilingual capabilities in large language models to enable cross-lingual collaborative reasoning in the financial domain remains insufficiently studied. To address this, we propose a **C**ross-**L**ingual **C**ollaboration (CLC) method: first, financial texts are aligned from one language to another based on semantic and syntactic structures, enabling the model to capture complementary linguistic features. Then, we integrate sentiment analysis results from both languages through redundancy removal and conflict resolution, enhancing the effectiveness of cross-lingual collaboration. Our experiments cover seven languages from three language families, including six UN official languages, and evaluate CLC on two English datasets and one Chinese dataset. Results show that multilingual collaboration improves sentiment analysis accuracy, especially among linguistically similar languages. Furthermore, stronger reasoning capabilities in LLMs amplify these benefits. Our code is available at https://anonymous.4open.science/r/Cross-lingual-Collaboration.

1 Introduction

In a dynamic and globalized financial environment, accurate entity-level sentiment analysis has become a cornerstone of various financial activities. It supports investors in making informed decisions, assists financial institutions in risk assessment, and enables companies to monitor market reputation. The rapid advancement of large-language models (LLMs), particularly their zero-shot capabilities, expands new opportunities for entity-level financial sentiment analysis (EFSA). Zero-shot learning enables models to analyze financial sentiment without explicit training on specific data, addressing the challenge of emerging financial entities and market trends. Meanwhile, the global financial landscape is inherently complex and multilingual, with each language reflecting a unique way of thinking (Vygotsky, 2012). While LLMs possess a natural advantage in processing financial texts across different languages (Li et al., 2023), how to leverage this capability for cross-lingual collaborative reasoning remains an open research question.

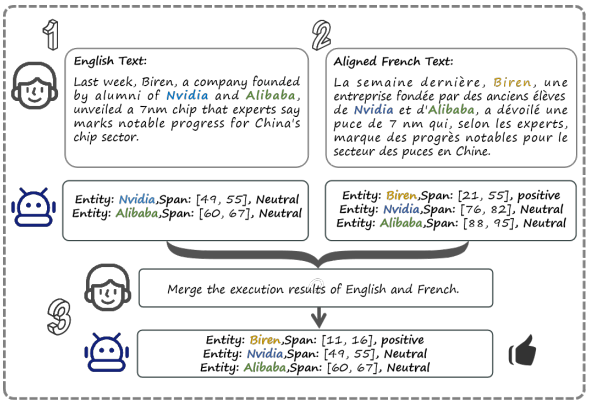


Figure 1: Our cross-lingual collaboration strategy in the financial text entity-level sentiment analysis task: an example in English and French. By integrating the execution results of both languages, better results are achieved in a zero-shot setting.

Existing research on entity-level financial sentiment analysis has several limitations. Most studies focus on single-language analysis, overlooking the benefits of multilingual information. Multilingual approaches often rely on simplistic integration methods, which may cause information loss and inaccuracies due to linguistic and cultural differences.

Our study proposes a cross-lingual collaboration approach. It first aligns financial texts between languages based on semantic and syntactic relationships, then integrates results to capture unique lin-

*Corresponding author

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2993–3003
December 20-24, 2025

guistic features and perspectives, improving EFSA performance (see Figure 1).

We examine language pairs from three language families to analyze intra- and inter-family interactions in sentiment analysis. Experiments use two English-language and one Chinese-language financial dataset.

The main contributions of this paper are as follows:

- We design a novel three-turn interactive zero-shot cross-lingual collaboration prompting method.
- Extensive experiments across multiple LLMs demonstrate that the CLC method enhances the zero-shot EFSA performance.
- Our research explores the performance differences among various languages and language families in cross-lingual collaboration, revealing the collaborative relationships among major languages.

2 Related Work

2.1 Entity-level Financial Sentiment Analysis

Natural language processing (NLP) techniques have been widely adopted in financial sentiment classification (Yang et al., 2022; Chuang and Yang, 2022; Xing et al., 2020). Existing financial sentiment classification datasets primarily provide annotations at the document or sentence level (Cortis et al., 2017; Huang et al., 2023a; Shah et al., 2023a). With the advancement of NLP models, sentiment analysis has evolved from coarse-grained to fine-grained approaches (Du et al., 2024). Recent studies have achieved state-of-the-art performance in benchmarks such as SemEval 2017 Task 5 and FiQA Task 1 (Du et al., 2023). The FiQA dataset introduces aspect-level sentiment annotation, but lacks entity-specific sentiment labels. For financial entity recognition, FiNER (Shah et al., 2023b) and FNXL (Sharma et al., 2023) provide entity annotations and numerical span recognition, respectively, but do not include sentiment annotations. The FinEntity dataset is currently one of the few datasets incorporating both financial entity spans and sentiment information (Tang et al., 2023b), where the span index precisely locates each entity mention to distinguish repeated entities with different sentiments. Zhu et al. (2025) construct a large-scale entity-level financial sentiment analysis dataset covering both English and Chinese[1], and enhance sentiment prediction accuracy through model calibration and example retrieval strategies.

[1] https://github.com/NLP-Bin/SILC-EFSA

On the other hand, the study by Chen et al. (2024) shows that large language models perform poorly on EFSA tasks under zero-shot settings when using the Chain-of-Thought (CoT) reasoning framework. This finding highlights that the complexity of financial entities far exceeds that of general text entities, thereby necessitating more advanced semantic understanding from models.

2.2 Cross-lingual Prompting

LLMs have demonstrated remarkable performance across a wide range of NLP tasks (Brown et al., 2020; Tang et al., 2023a; Qin et al., 2024). Unlike traditional Pre-trained Language Models (PLMs) (Wei et al., 2022; Wang et al., 2022), LLMs enable zero-shot learning without requiring modifications to model parameters during training and inference, making them highly versatile (Wei et al., 2022; Feng et al., 2023; Zhang et al., 2023). Shi et al. (2022) introduce the first multilingual dataset to assess the mathematical reasoning abilities of LLMs, laying the foundation for research in cross-lingual CoT. Prior research highlights the effectiveness of LLMs in various cross-lingual tasks (Chai et al., 2024; Huang et al., 2023b; Tanwar et al., 2023), including spoken language understanding and summarization. However, Qin et al. (2023) introduce Cross-lingual Prompting (CLP), a novel zero-shot approach that aligns CoT reasoning across languages without requiring additional training data. Furthermore, they propose Cross-lingual Self-consistent Prompting (CLSP), leveraging structured multilingual reasoning pathways designed by linguistic experts to enhance model consistency and performance. Zhang et al. (2024) further design an automated language selection and weight allocation, achieving superior performance. Chen et al. (2024) demonstrate that entity-level financial sentiment analysis performs poorly in a single-language inference setting. With over 200 countries and 7,000 languages worldwide, multilingual inference presents the potential for performance improvement.

3 Methodology

As shown in Figure 2, we propose a novel cross-lingual collaboration method for enhancing zero-shot EFSA, which is described in detail below.

3.1 Language Text Alignment

To achieve cross-lingual alignment, we define a translation function $\mathcal{T}$ implemented by a large lan-

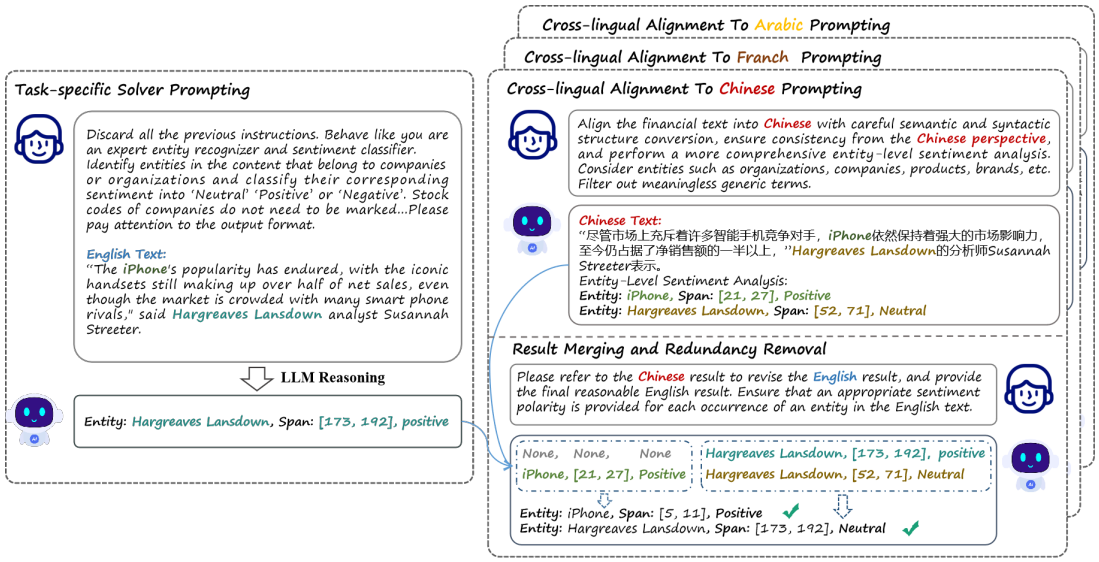


Figure 2: Overview of our cross-lingual collaboration method. The CLC method involves a three-turn interactive process, consisting of: instructions for executing the EFSA-specific task, cross-lingual alignment instructions from the source language to the collaborative language, and finally, instructions for fusing the results from both languages.

guage model $\mathcal{M}$. The translation process is formulated as follows.

$$\mathcal{T}(S) = \arg\max_{T} \prod_{j=1}^{m} \mathbb{P}(y_j \mid y_{<j}, S, \theta_{\mathcal{M}}) \quad (1)$$

$$T = \mathcal{T}(S) \quad (2)$$

where S and T denote the source and target language financial texts, respectively. The model generates T by maximizing the conditional probability of each token y_j, given the previously generated tokens $y_{<j}$ and the source text S, ensuring semantic alignment.

3.2 Cross-lingual Entity-Level Sentiment Analysis

After alignment, we perform financial sentiment analysis on both S and T. We define the sentiment classification function as:

$$\mathcal{F} : \{S, T\} \rightarrow \mathbf{A} \cup \mathbf{A}' \quad (3)$$

$$\mathbf{A} \cup \mathbf{A}' = \{(e_k, p_k) \mid e_k \in \{S, T\}\} \quad (4)$$

$$\mathcal{M}(\{S, T\}) = \mathcal{F}(\{S, T\}) \quad (5)$$

where e_k represents detected financial entities from either S or T, and p_k denotes their corresponding sentiment polarity. The large language model $\mathcal{M}$ applies the function $\mathcal{F}$ to conduct entity-level sentiment analysis across languages, ensuring semantic consistency.

3.3 Result Merging and Redundancy Removal

3.3.1 Entity and Sentiment Merging

After obtaining entity-level sentiment predictions from the large language model $\mathcal{M}$, we integrate results from both languages using the merging function:

$$\mathcal{M}_c(\mathbf{A}, \mathbf{A}') = \mathbf{C} = \{(e_k, p_k) \mid e_k \in S \cup T\} \quad (6)$$

where $\mathbf{A} \cup \mathbf{A}'$ represents the entity-level sentiment predictions from S and T.

3.3.2 Redundancy Removal and Conflict Resolution

To ensure consistency, we apply a redundancy removal function $\mathcal{R}$, which consolidates duplicate entities:

$$\mathcal{R}(\mathbf{C}) = \{(\tilde{e}_q, \tilde{p}_q) \mid \tilde{e}_q \text{ is unique}\} \quad (7)$$

If sentiment conflicts arise for the same entity across languages, we resolve them using a probabilistic conflict resolution function:

$$\mathcal{H}(\tilde{p}_q) = \arg\max_{p} \mathbb{P}(p \mid S, T, \theta_{\mathcal{M}}) \quad (8)$$

where $\mathbb{P}(p \mid S, T, \theta_{\mathcal{M}})$ represents the model's estimated probability distribution over sentiment polarities. The final sentiment predictions are:

$$\mathbf{F} = \mathcal{H}(\mathcal{R}(\mathbf{C})) \quad (9)$$

	FinEntity	SEntFiN-Span	FinEntCN
Number of Texts	979	10753	10832
Single Entity Texts	390 (39.84%)	7897 (73.44%)	8194 (75.65%)
Multiple Entity Texts	589 (60.16%)	2856 (26.56%)	2638 (24.35%)
Average Text Length by Tokens	37.01	9.91	145.23
Positive Entities	503 (23.60%)	5084 (35.21%)	8037 (53.89%)
Negative Entities	498 (23.37%)	3828 (26.51%)	5040 (33.79%)
Neutral Entities	1130 (53.03%)	5527 (38.28%)	1838 (12.32%)
All Entities	2131	14439	14915
Average Entity Num Per Text	2.18	1.34	1.38

Table 1: The statistics of entity-level financial sentiment analysis datasets.

This approach ensures optimal cross-lingual sentiment integration, enhancing the accuracy of financial sentiment analysis.

4 Experiments

4.1 Datasets

Our experimental datasets include the FinEntity dataset constructed by Tang et al. (2023b) and the dataset created by Zhu et al. (2025), which comprises two subsets: SEntFiN-Span and FinEntCN. This enables our experiments to be conducted on the most comprehensive dataset to date, covering both English and Chinese financial texts. The statistical information of the dataset is presented in Table 1. In each experiment, 200 samples are selected from the test sets of the three datasets. Although the sample size is small, preliminary experiments show stable performance, and each experiment is repeated three times to ensure robustness.

4.2 Implementation Settings

In our experiments, we select GPT-3.5[2], GPT-4o, DeepSeek-V2-Chat[3], and qwen-turbo[4] as four advanced large language models to serve as our experimental models. The experimental process consists of language alignment, task execution, and result fusion, all of which are independently performed by the same model to ensure consistency and controllability. We adopt a strict scoring system, where a test example is considered correctly classified only when all named entities and their boundaries in the example are accurately identified and their associated sentiments are correctly classified.

For fair comparison, top-p and temperature are fixed to 1.0 and 0.0, respectively, in all experiments. Additionally, we focus on seven languages from three major language families with large user populations (Indo-European, Sino-Tibetan, and Afro-Asiatic), including six official United Nations languages. We systematically investigate their impact on English and Chinese task collaboration.

4.3 Main Results

The experimental results are presented in Table 2. Experimental results indicate that cross-lingual collaboration significantly improves sentiment analysis accuracy in most cases. Meanwhile, we observe that major languages generally enhance performance in cross-lingual collaboration, while minor languages may have the opposite effect. In language combination experiments, collaborations within the same language family yield the most notable improvements. For instance, the combinations of English with French, Spanish, and Russian, as well as Chinese with Tibetan, enhance sentiment analysis accuracy. Additionally, English-Chinese collaboration demonstrates widespread performance gains. However, cross-family combinations (e.g., English with Sino-Tibetan or Afro-Asiatic languages) may lead to performance degradation, suggesting that linguistic structure and grammatical differences significantly impact cross-lingual collaboration.

5 Experimental Analysis

5.1 Impact of CLC on Model Performance

According to the experimental data, cross-lingual collaboration (CLC) contributes to improved model performance. As shown in Figure 3, using the GPT-4o model as an example, the F1 score improves substantially when collaborating with most languages. Notably, Chinese and English achieve the highest overall improvement when collaborating with

[2] https://openai.com
[3] https://chat.deepseek.com
[4] https://huggingface.co/Qwen

Method	Model	FinEntity(English)			SEntFiN-Span(English)			FinEntCN(Chinese)		
		P	R	F1	P	R	F1	P	R	F1
Origin	G-3.5	0.5253	0.6723	0.5902	0.5143	0.6923	0.5902	0.1564	0.3725	0.2203
	G-4o	0.7989	0.6714	0.7296	0.5354	0.6721	0.5952	0.2097	0.4468	0.2854
	DS-V2	0.5978	0.6370	0.6168	0.5839	0.6192	0.6010	0.2276	0.3972	0.2894
	qw-tb	0.3115	0.2670	0.2875	0.2939	0.3096	0.3016	0.1756	0.2660	0.2116
Indo-European Family										
French-Co	G-3.5	0.5619	0.7143	0.6289	0.6850	0.5766	0.6261	0.2329	0.3420	0.2771
	G-4o	0.8021	0.7624	0.7817	0.6934	0.5920	0.6387	0.2604	0.3846	0.3106
	DS-V2	0.6481	0.6451	0.6465	0.6124	0.6690	0.6395	0.2410	0.4270	0.3081
	qw-tb	0.3844	0.3466	0.3645	0.2517	0.2598	0.2557	0.1704	0.2979	0.2168
Spanish-Co	G-3.5	0.6552	0.6032	0.6281	0.5293	0.6215	0.5717	0.1694	0.3696	0.2323
	G-4o	0.8690	0.7812	**0.8203**	0.6257	0.6006	0.6129	0.2788	0.4565	0.3462
	DS-V2	0.6358	0.6745	0.6545	0.5990	0.6500	0.6235	0.2432	0.4091	0.3051
	qw-tb	0.4143	0.3625	0.3867	0.2457	0.2562	0.2509	0.1940	0.3191	0.2413
Russian-Co	G-3.5	0.6054	0.6792	0.6402	0.5517	0.6957	0.6154	0.1652	0.4043	0.2346
	G-4o	0.7963	0.7771	0.7866	0.5827	0.6539	0.6163	0.2761	0.4054	0.3285
	DS-V2	0.6231	0.6581	0.6401	0.6006	0.6904	0.6424	0.2704	0.4823	0.3465
	qw-tb	0.3818	0.3443	0.3621	0.2876	0.3950	0.3328	0.1876	0.3227	0.2373
English-Co	G-3.5	-	-	-	-	-	-	0.1652	0.4043	0.2346
	G-4o	-	-	-	-	-	-	0.2761	0.4054	0.3285
	DS-V2	-	-	-	-	-	-	0.2704	0.4823	0.3465
	qw-tb	-	-	-	-	-	-	0.1876	0.3227	0.2373
Sino-Tibetan Family										
Chinese-Co	G-3.5	0.5364	0.6909	0.6039	0.5148	0.6758	0.5784	-	-	-
	G-4o	0.8021	0.7424	0.7711	0.6263	0.5939	0.6097	-	-	-
	DS-V2	0.6404	0.5854	0.6621	0.6311	0.7196	**0.6725**	-	-	-
	qw-tb	0.4715	0.4075	0.4372	0.3055	0.3381	0.3209	-	-	-
Tibetan-Co	G-3.5	0.5286	0.5747	0.5507	0.5833	0.5000	0.5385	0.2040	0.3723	0.2493
	G-4o	0.7108	0.6614	0.6852	0.6047	0.4943	0.5440	0.3690	0.5586	**0.4444**
	DS-V2	0.7450	0.7136	0.7181	0.5672	0.6470	0.6048	0.2899	0.4348	0.3478
	qw-tb	0.3882	0.2435	0.2993	0.2548	0.1886	0.2168	0.2440	0.2872	0.2638
Afro-Asiatic Family										
Arabic-Co	G-3.5	0.5091	0.6585	0.5692	0.5238	0.7097	0.6027	0.1839	0.4291	0.2574
	G-4o	0.8039	0.6508	0.7193	0.6252	0.5981	0.6113	0.3118	0.4775	0.3772
	DS-V2	0.6812	0.7072	0.6847	0.5942	0.6643	0.6273	0.3110	0.4380	0.3637
	qw-tb	0.3259	0.2803	0.3014	0.2563	0.2527	0.2545	0.1970	0.3250	0.2453

Table 2: Experimental results on two English datasets (FinEntity, SEntFiN-Span) and one Chinese dataset (FinEntCN). "Origin" refers to directly performing the EFSA task in the source language. "French-Co" represents collaboration between the source language and French, with other languages following a similar notation. "G-3.5" denotes the GPT-3.5 model, while "G-4o" denotes the GPT-4o model. "DS-V2" refers to DeepSeek-V2-Chat, and "qw-tb" represents qwen-turbo. The best performance of each model is underlined, and the best performance on each dataset is shown in bold. We evaluate the experimental results using Precision (P), Recall (R), and macro-F1 (F1), and report the average over three independent runs.

Spanish, with an average increase of 5.64 percentage points. The other three models also exhibit consistent performance gains across most language collaborations. These results suggest that the CLC method enables models to better comprehend and reason about sentiment information.

5.2 Wider Languages Perform Better

The generalizability of the CLC method is validated across multiple datasets and languages. Based on common sense, we hypothesize that the performance of large language models is highly correlated with the proportion of pretraining data avail-

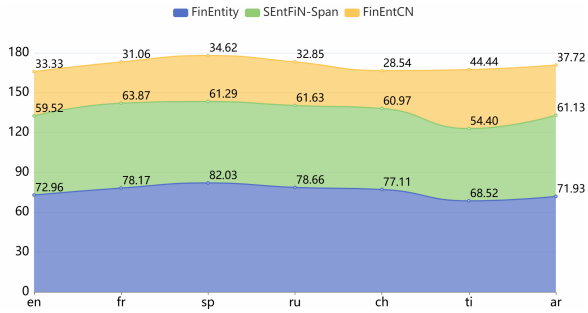


Figure 3: The F1 score performance of the GPT-4o model across three datasets.

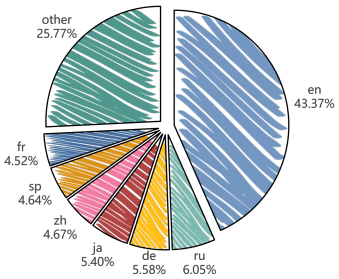


Figure 4: The language distribution of Common Crawl in 2024.

able for each language(Blevins and Zettlemoyer, 2022; Malkin et al., 2022). To investigate this, we analyze the language distribution in the widely used multilingual pretraining dataset, Common Crawl 2024[5] (see Figure 4). For some low-resource languages, such as Tibetan, the effectiveness of the CLC method is relatively weaker and may even introduce noise. For instance, in GPT-series models, collaboration between English and Tibetan resulted in an approximately 4.5 percentage point decrease in performance. However, the DeepSeek model achieved performance gains when English and Chinese collaborated with Tibetan. This suggests that the generalizability of cross-lingual prompting depends on language resource availability.

5.3 Three-turn Prompting is Better Than Single-turn Prompting

In our experiments, this three-turn approach significantly enhances the model's performance when compared to single-turn prompting. As shown in Figure 5, taking the FinEntity dataset as an example, GPT-4o achieves a higher F1 score with the three-turn prompt, on average improving by 2.55%. This indicates that the three-stage interactive prompting helps the model gain sufficient contextual understanding in the initial phase, followed by more refined reasoning in subsequent steps, resulting in more accurate sentiment analysis.

[5] https://www.commoncrawl.org/

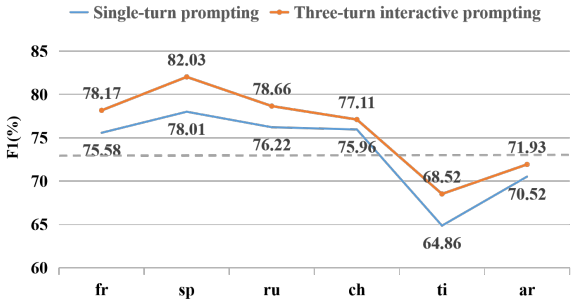


Figure 5: F1 Score Comparison on the FinEntity: Three-turn Interactive Prompting vs. Single-turn Prompting (GPT-4o).

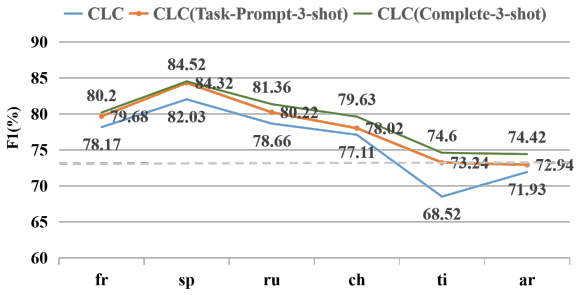


Figure 6: The performance of GPT-4o on the FinEntity dataset with few-shot learning is shown.

5.4 CLC Adapts to Few-shot Setting

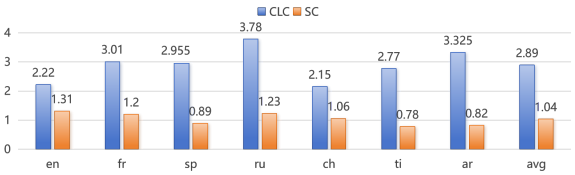


Figure 7: The table presents the experimental results of the average F1 score improvement across three datasets using four large language models, comparing Cross-Lingual Collaboration (CLC) with standard Self-Consistency (SC).

Few-shot learning further enhances the effectiveness of the CLC method. On the FinEntCN dataset, after incorporating task examples through few-shot learning, GPT-4o shows an average F1 score improvement of 2%. When three-stage prompting also utilizes few-shot learning, the F1 score increases by approximately 3% (see Figure 6). "Task-Prompt" refers to incorporating examples in the

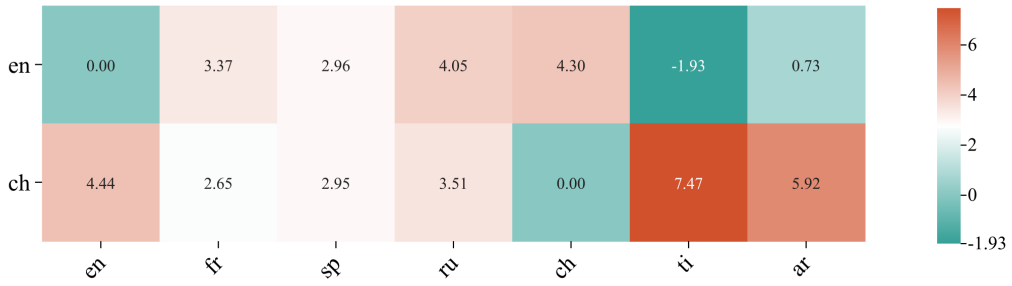


Figure 8: The collaboration results of English and Chinese with other languages. Red indicates an improvement in F1 score, while green indicates a decrease in F1 score.

first round of specific task prompts, while "Complete" indicates the inclusion of examples in all three rounds of prompting. This demonstrates the adaptability of the CLC method and its effective support for sentiment analysis tasks.

5.5 Cross-lingual Self-consistent Prompting Surpasses Vanilla Self-consistency

To verify the effectiveness of CLC, we conduct a vanilla self-consistency (VSC) experiment(Wang et al., 2022). As shown in Figure 7, the experimental results show that across three datasets, multiple models using the CLC method achieve an F1 score 1.85 percentage points higher than standard self-consistency prompting. This improvement indicates that, compared to a monolingual setting, cross-lingual prompting more effectively integrates multilingual perspectives and reduces model bias.

5.6 Different Language Family Bring Different Effects

As shown in Figure 8, from a linguistic family perspective, collaboration between English and other Indo-European languages, such as French, Spanish, and Russian, consistently enhances performance, with an average F1 score increase of 3.46%. Collaboration within the Sino-Tibetan family produces an average improvement of 1.18%, while cooperation within the Afro-Asiatic family shows minimal gains. Notably, collaboration between Chinese and Tibetan, both part of the Sino-Tibetan family, leads to a substantial enhancement, with an F1 score increase of 7.47%, while collaboration with Indo-European languages results in an average improvement of 3.39%.

5.7 Linguistic Syntactic Analysis

To systematically analyze the cross-lingual interaction effects in multilingual financial sentiment analysis, we conduct a multidimensional linguistic evaluation across seven languages—English, Chinese, Spanish, French, Russian, Arabic, and Tibetan. This evaluation encompasses syntactic structure, information density, and sentiment lexicon mapping consistency.

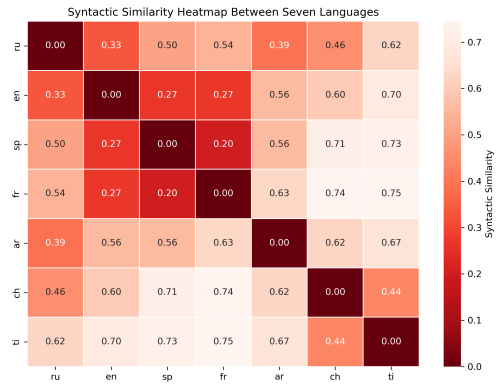


Figure 9: Syntactic similarity heatmap among seven languages based on POS trigram Jensen-Shannon distances. Darker red indicates higher similarity (lower distance).

Following the methodology proposed by (De Gregorio et al., 2024), we construct a syntactic similarity matrix based on Jensen-Shannon distances computed over POS trigram distributions derived from financial text corpora. The heatmap in Figure 9 visualizes the relative syntactic distances among the seven languages. This approach captures differences in surface word order and syntactic alignment, reflecting the degree of structural similarity. The results reveal clear clustering effects along language family lines. For instance, Indo-European languages such as English, French, Spanish, and Russian exhibit higher syntactic similarity, as they share many grammatical features (e.g., SVO word order and similar morphological structures[6]). In contrast, Tibetan shows substantial

[6] https://wals.info

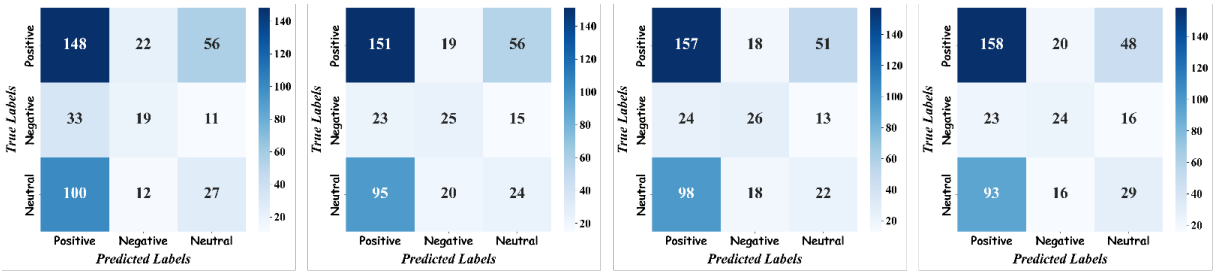


Figure 10: Confusion Matrix Analysis of DeepSeek-V2 on the FinEntity Dataset for four settings: (1) inference with the original language, (2) collaborative inference with Spanish, Chinese, Arabic.

syntactic divergence due to its agglutinative morphology and post-verbal predicate order. Chinese, as a prototypical isolating language, exhibits a flatter syntactic dependency distribution and higher information density.

Table 3: Linguistic Dimension Comparison (Relative to English). **Info. Dens**: Information Density; **Term Cons.**: Terminology Consistency, reflecting the stability and uniformity of specialized vocabulary; **Sent. Align.**: Sentiment Polarity Alignment.

Language	Info. Dens. (bits/token)	Term Cons.	Sent. Align.
en	4.03	1.00	1.00
zh	9.71	0.88	0.87
sp	4.01	0.94	0.91
fr	3.98	0.91	0.89
ru	4.35	0.87	0.86
ar	4.31	0.82	0.81
ti	4.00	0.78	0.79

As shown in Table 3, the "Sentiment Lexicon Mapping" column measures the translation accuracy and polarity consistency of core financial sentiment terms such as "crisis," "volatility," and "bullish" across different languages. This metric is computed using aligned bilingual term pairs from dictionaries and parallel financial corpora, and evaluates their polarity preservation. Language pairs with similar syntactic structures and shared financial expression conventions (e.g., English and Spanish) tend to demonstrate higher mapping consistency. In contrast, languages such as Arabic and Tibetan may exhibit lower mapping accuracy due to lexical gaps or sentiment ambiguity. Structural compatibility—particularly in syntax and sentiment expression—plays a crucial role in the performance of zero-shot cross-lingual collaboration. When structural disparities are large, models require more adaptive prompting strategies to avoid misleading sentiment inference. Models are generally adept at enhancing performance through agreement, but less effective at resolving ambiguity.

5.8 CLC Quantitative Analysis

As shown in Figure 10, it is evident that CLC positively influences the model's sentiment classification capability. The DeepSeek-V2 model shows an overall improvement in accuracy when performing inference with Spanish, Chinese, and Arabic languages, demonstrating strong cross-lingual integration ability. Notably, the number of correct predictions in the Positive category increases, while the classification accuracy of the Negative category is also enhanced.

6 Conclusion

This study investigates the role of cross-lingual collaboration in zero-shot entity-level sentiment analysis in the financial domain and proposes the CLC method. Experimental results demonstrate that CLC can significantly enhance sentiment analysis performance by leveraging collaboration across different languages. Further analysis reveals several key findings: First, the effectiveness of collaboration is closely related to the availability of training data, with better performance observed when collaborating between resource-rich languages, whereas collaboration with low-resource languages does not always yield stable improvements. Second, language models with stronger reasoning capabilities exhibit greater stability in cross-lingual collaboration, and multi-turn iterative reasoning outperforms single-turn reasoning. Finally, collaboration among languages within the same language family tends to result in more substantial performance gains. We believe these findings contribute to future research on expanding cross-lingual collaboration to more languages and optimizing fusion strategies.

Limitations

This study has several limitations. First, the experiments involve only the major languages from the primary language families, and the results may not be generalizable to all languages. The fusion strategy employed may not be the optimal solution for all language combinations. Second, the performance of the method may vary depending on the model size and specific language capabilities. Lastly, the study focuses on financial sentiment analysis and does not fully explore the applicability of the method in other domains.

Acknowledgments

The authors thank the anonymous reviewers for their insightful comments. This work is mainly supported by the Key Program of the Natural Science Foundation of China (NSFC) (No.U23A20316).

References

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of english pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *arXiv preprint arXiv:2401.07037*.

Tianyu Chen, Yiming Zhang, Guoxin Yu, Dapeng Zhang, Li Zeng, Qing He, and Xiang Ao. 2024. EFSA: Towards event-level financial sentiment analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7455–7467, Bangkok, Thailand. Association for Computational Linguistics.

Chengyu Chuang and Yi Yang. 2022. Buy tesla, sell ford: Assessing implicit stock market preference in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 100–105.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 519–535.

Juan De Gregorio, Raúl Toral, and David Sánchez. 2024. Exploring language relations through syntactic distances and geographic proximity. *EPJ Data Science*, 13(1):61.

Kelvin Du, Frank Xing, and Erik Cambria. 2023. Incorporating multiple knowledge sources for targeted aspect-based financial sentiment analysis. *ACM Transactions on Management Information Systems*, 14(3):1–24.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9):1–42.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: a theoretical perspective. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 70757–70798.

Allen H Huang, Hui Wang, and Yi Yang. 2023a. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023b. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023a. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679.

Agam Shah, Ruchit Vithani, Abhinav Gullapalli, and Sudheer Chava. 2023b. Finer: Financial named entity recognition dataset and weak-supervision model. *arXiv preprint arXiv:2302.11157*.

Soumya Sharma, Subhendu Khatuya, Manjunath Hegde, Afreen Shaikh, Koustuv Dasgupta, Pawan Goyal, and Niloy Ganguly. 2023. Financial numeric extreme labelling: A dataset and benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3550–3561.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023a. Large language models are in-context semantic reasoners rather than symbolic reasoners. *arXiv preprint arXiv:2305.14825*.

Yixuan Tang, Yi Yang, Allen Huang, Andy Tam, and Justin Tang. 2023b. FinEntity: Entity-level sentiment classification for financial texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15465–15471, Singapore. Association for Computational Linguistics.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. *arXiv preprint arXiv:2305.05940*.

Lev S Vygotsky. 2012. *Thought and language*, volume 29. MIT press.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 24824–24837.

Frank Xing, Lorenzo Malandri, Yue Zhang, and Erik Cambria. 2020. Financial sentiment analysis: an investigation into common mistakes and silver bullets. In *Proceedings of the 28th international conference on computational linguistics*, pages 978–987.

Yi Yang, Kunpeng Zhang, and Yangyang Fan. 2022. Analyzing firm reports for volatility prediction: A knowledge-driven text-embedding approach. *INFORMS Journal on Computing*, 34(1):522–540.

Yongheng Zhang, Qiguang Chen, Min Li, Wanxiang Che, and Libo Qin. 2024. AutoCAP: Towards automatic cross-lingual alignment planning for zero-shot chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9191–9200, Bangkok, Thailand. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Senbin Zhu, ChenYuan He, Hongde Liu, Pengcheng Dong, Hanjie Zhao, Yuchen Yan, Yuxiang Jia, Hongying Zan, and Min Peng. 2025. Silc-efsa: Self-aware in-context learning correction for entity-level financial sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4980–4992.

A Appendix

The instruction design of our cross-lingual collaboration strategy is as follows.

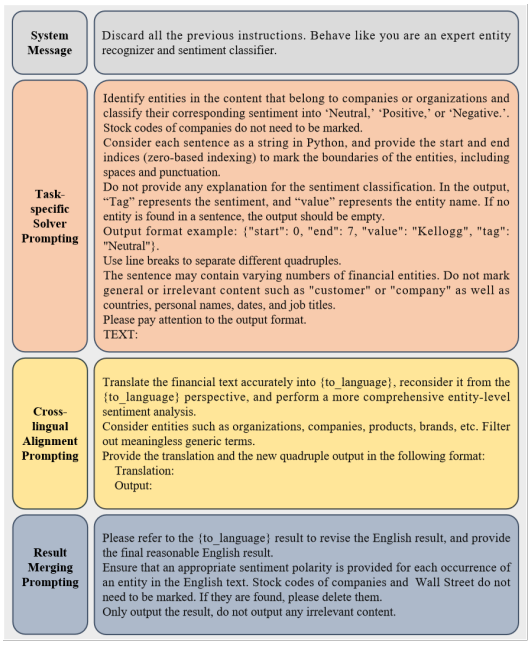
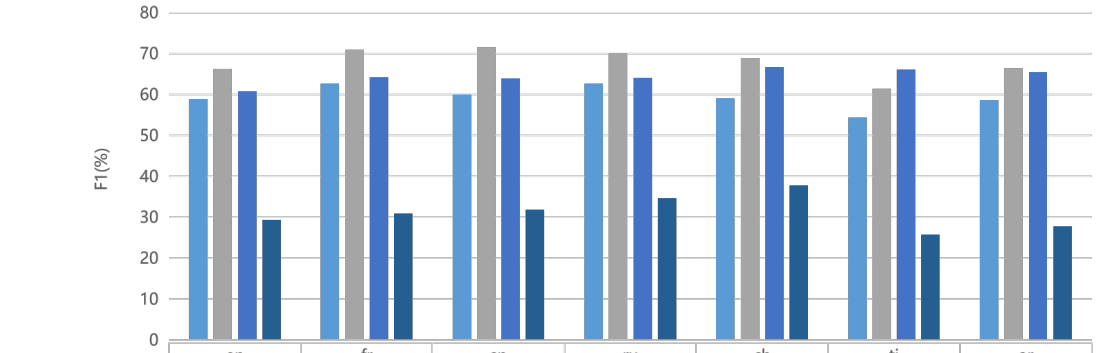


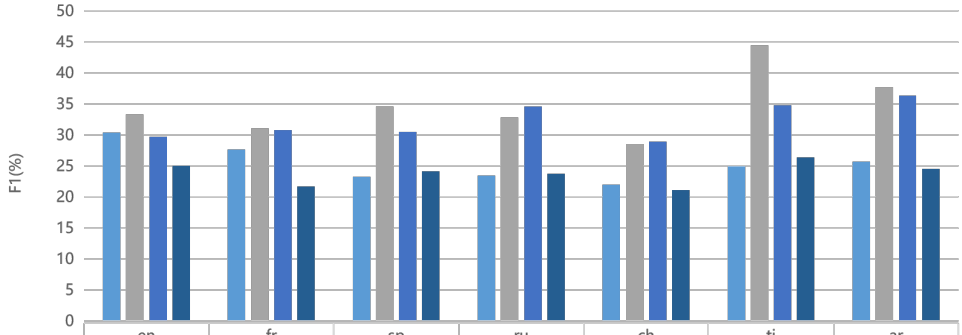
Figure 11: The English instruction template of our CLC.

The CLC performance of different LLMs is shown in Figure 12 and Figure 13.



	en	fr	sp	ru	ch	ti	ar
GPT-3.5	59.02	62.75	59.99	62.78	59.115	54.46	58.595
GPT-4o	66.24	71.02	71.66	70.145	69.04	61.46	66.53
DeepSeek-V2-Chat	60.89	64.3	63.9	64.125	66.73	66.145	65.6
qwen-turbo	29.46	31.01	31.88	34.745	37.905	25.805	27.795

Figure 12: The average F1 scores of different models on two English datasets.



	en	fr	sp	ru	ch	ti	ar
GPT-3.5	30.42	27.71	23.23	23.46	22.03	24.93	25.74
GPT-4o	33.33	31.06	34.62	32.85	28.54	44.44	37.72
DeepSeek-V2-Chat	29.68	30.81	30.51	34.65	28.94	34.78	36.37
qwen-turbo	25	21.68	24.13	23.73	21.16	26.38	24.53

Figure 13: The F1 scores of different models on the Chinese dataset.