AMR-GCC: Two-step Cross-Document Event Factuality Identification on Data Augmentation

Zijie Qian, Zhong Qian(✉), Chengwei Liu, and Peifeng Li

School of Computer Science and Technology, Soochow University, Suzhou, China
20225227021@stu.suda.edu.cn, qianzhong@suda.edu.cn,
20194027010@stu.suda.edu.cn, pfli@suda.edu.cn

Abstract. This paper focuses on Cross-Document Event Factuality Identification (CEFI) which aims to infer the factual accuracy of a cross-document event from several different document-level texts. As a latest task in Event Factuality Identification (EFI) and an expanded task of Document-level Event Factuality (DEFI), CEFI does not have any corpus or methods, since existing related work is limited to document-level task. Based on these issues, we construct a new corpus, which faces the sparsity of the data and the uneven distribution with different labels limit due to the characteristics of news corpus, and propose Abstract Meaning Representation Graph-based Cross-Document Classification (AMR-GCC) as a two-step method. Combined with additional data generated by fine-tuning GLM-32B in terms of data augmentation, we divide AMR-GCC into two sub tasks to improve the effectiveness of long text processing, i.e., document-level event classification and event collaboration. The results show our method outperforms several SOTAs.

Keywords: Cross-Document Event Factuality Identification · LLM-Powered Data Augmentation · Abstract Meaning Representation.

1 Introduction

This paper focuses on Cross-Document Event Factuality Identification (CEFI), which is defined as assessing the level of factual accuracy in cross-document events. That is, determining whether something in a text is a fact, a possibility, and their polarity: positive or negative. Event factuality is categorized into five distinct types as follows [13]: CerTain Positive, PoSsible Positive, CerTain Negative, PoSsible Negative, Underspecified. CEFI differs from other Event Factuality Identification (EFI) tasks in terms of text categories by using multiple text unions. Previous EFI encompasses two main components: Sentence-level Event Factuality Identification (SEFI) [12] and Document-level Event Factuality Identification (DEFI) [13], which involve establishing event factuality by analyzing either individual sentences or entire documents. In addition, CEFI can be seen as an extension task of DEFI and the latest task in EFI.

As illustrated by Fig. 1, we provide a simplified example. We extract the main sentences from five documents, each with its own understanding of the cross-document event factuality. The facts stated by D2 and D4 show the phenomenon of economic recession, and their labels are CT+. "low levels of layoffs" in D1 and "an incredibly high demand for workers" in D5 indicate that the economic recession does not occur,

Cross-document Event: The US labor market is in a recession.	
Cross-Document Event Factuality: CT-	
[D1:CT-]	The latest Job Openings and Labor Turnover Survey showed there were 11.3 million job openings in May, or 1.9 positions for every job seeker, and historically low levels of layoffs.
[D2:CT+]	While that's good news for job seekers, there are also signs that employers are starting to cut back.
[D3:CT-]	The US labor market is clearly not in a recession. Unemployment can be a indicator.
[D4:CT+]	"We are already seeing some companies start to pull back on hiring, whether through hiring freezes or just quite simply acting slower to fill their existing job openings," said Daniel Zhao, senior economist for Glassdoor.
[D5:CT-]	More broadly, the US labor market is still in a period marked by an incredibly high demand for workers, said Layla O'Kane, senior economist for labor market data firm Lightcast.

Fig. 1. An example of cross-document event factuality identification. The colored words are important information used to make judgments.

therefore their labels are CT-. "Clearly not in a recession" in D3 directly indicates the impossibility of an economic recession. The persuasiveness of the CT+ label here is not enough, resulting in the cross-document event factuality being CT-.

In the past, significant advancements have been made in the field of DEFI. Qian et al. [13] utilized a BiLSTM model to encode both sentences and syntactic paths, complemented by an attention mechanism to seamlessly integrate semantic and syntactic information. Cao et al. [1] introduced an innovative uncertain local-to-global network, effectively merging local uncertainty with global structural insights, thereby achieving state-of-the-art performance. Furthermore, Zhang et al. [16] pioneered a novel evidence sentence selection task and developed a pipeline method that decoupled evidence sentence selection from event fact recognition, also yielding significant success. However, previous research did not pay attention to the situation where multiple chapters have the same main event, and this situation exists very frequently in reality, which is a drawback. Therefore, we introduce Cross-Document into EFI.

The concept of Cross-Document has been used in many NLP tasks, such as sentiment analysis and question answering. CEFI aims to gather events from multiple documents together to determine the cross-document event factuality, which focuses on the relationship between various documents and the cross-document event. This task can offer a more comprehensive understanding of information, improving the accuracy of fact-checking, and enhancing event tracking and monitoring. For example, determining the authenticity of information from multiple sources to assist decision-making, banning false information, or helping manage. Simultaneously, we introduce an additional category: NA, which represents the entire document being unrelated to the cross-document event.

Based on the lack of previous research on CEFI, we learn from the experience of DEFI and consider that there are two difficulties in the current research on CEFI: 1)

There are almost no available datasets. The definition of CEFI determines the size of its dataset, which requires a lot of time and manpower. Meanwhile, the annotation of the CEFI corpus has a similar issue to the other EFI tasks, the sparsity of the data and the uneven distribution of corpora with different labels. 2) Although research on CEFI can draw on DEFI, there is still a need for improvement due to significant structural differences. In contrast to DEFI, CEFI incorporates additional relationships between multiple documents and the cross-document event. However, the associations between different documents are weak. Additionally, the inherent challenge of encoding long text tasks is unavoidable.

With the emergence of fine-tuning techniques for large language models, there are many examples of applying them to low resource problems, such as using fine-tuning LLM to solve the problem of few-shot. Based on this, on the corpus issue on CEFI, we fine-tune LLM to generate few-shot data after annotating a dataset. Meanwhile, we validate the data generated by fine-tuning LLM and demonstrate its effectiveness, which has similar features to the original data.

For long text processing, current methods are more inclined towards extracting multi-granularity information [3]. Inspired by this and combined with CEFI's definition, We propose AMR-GCC, a new method, which divides CEFI into two sub tasks. It focuses on extracting comprehensive features that align with the respective labels and integrating them to derive holistic full-text features.

Overall, our contributions can be summarized by the following four points:

- We introduce the concept of Cross-Document into Event Factuality Identification for the first time, which is a major innovation in EFI tasks.
- We annotate a new dataset based on CEFI's definition, which has 4294 samples of data, and attempt to compensate for the deficiency by fine-tuning LLM. In order to balance reproducibility and performance, we choose GLM-32B to generate new data that conforms to the annotation style.
- We propose AMR-GCC, a new method, which divides CEFI into two sub tasks based on this characteristic: document-level event classification and event collaboration. It makes the text features more prominent to solve the problem of low accuracy in long text encoding, improving the robustness of the model and enhancing its performance.
- We put forward a practical method, the combination of fine-tuning LLM to obtain more semantic features and classification methods greatly improves accuracy, which also enhances the robustness and scalability of the method.

2 Related Work

The research on EFI started very early, and many models have been applied to SEFI and have demonstrated excellent performance. Murayama et al. [9] provided a detailed description of some datasets and their characteristics on EFI tasks and summarizes many previous related methods. In contrast, DEFI started slightly later but has also conducted sufficient research beforehand. Qian et al. [13] developed a bilingual corpus comprising Chinese and English documents and introduced an LSTM neural network augmented with adversarial training, which incorporated both intra-sequence and inter-sequence attention mechanisms to accurately identify document-level event factuality. Zhang et al. [17] expanded the scope of their DEFI research by integrating the nuanced aspects of negation and speculation, by one comprehensive approach allowing for a more thorough analysis of the subtleties within the text, which captured the intricacies of negative and speculative expressions. Cao et al. [1] introduced a groundbreaking uncertain local-to-global network, which effectively integrates local uncertainty with overarching global structural information. Additionally, Zhang et al. [16] introduced a novel evidence sentence selection task and devised a pipeline method that treated evidence sentence selection and event fact recognition as separate tasks, also achieving notable success.

The concept of cross-document was proposed early and has been applied to multiple related tasks in NLP. Nowadays, research based on it has also made significant progress. Langhe et al. [7] proposed a methodology combining traditional mention-pair coreference models with a lightweight and modular graph reconstruction algorithm on Cross-Document Event Coreference. Zhang et al. [18] focused on improving Cross-Document Summarization by combining the use of auxiliary entity and event recognition systems with incorporating an alignment loss between IE nodes and their text spans.

Existing research in DEFI has not fully addressed the complexities and nuances inherent in document-level event factuality. There are notable gaps in comprehensive methodologies, robust datasets, and advanced modeling techniques. Additionally, cross-document processing faces notable challenges, including the intricacies of long text handling, the complexity of integrating disparate information across multiple documents, and the compounded difficulties arising from these integration processes.

3 Method

In this section, We introduce the specific task definition of CEFI (Sect. 3.1) and outline all the steps of our method (Sect. 3.2). Firstly, we fine-tune the large model to generate data (Sect. 3.3), and then verify the validity of the data (Sect. 3.4). Finally, we introduce the two components of our two-step approach AMR-GCC: document-level event classification and event collaboration (Sect. 3.5).

3.1 Problem Statement

The definition of CEFI is to identify the cross-document event factuality in multiple joint documents. We can use information obtained from these texts to determine the cross-document event factuality. The factuality values can be divided into the following five categories [13]: CerTain Positive (CT+), PoSsible Positive (PS+), CerTain Negative (CT-), PoSsible Negative (PS-), Underspecified (Uu). Additionally, there is a value within a separate document that signifies the event factuality specific to that document. This value encompasses six possibilities: the five previously mentioned values, along with “NA”, which indicates that the document is unrelated to the cross-document event and can be deemed as irrelevant or useless information.

3.2 Overall Sub-modules

As illustrated by Fig. 2, the methodology of this paper is generally divided into three parts, data generation, data validation, and event factuality identification, improving CEFI performance from both data and model perspectives.

Data Generation: We generate few-shot data and use it as fine-tuning corpus input into LLM, generating similar titles and content to ensure the similarity and usability of the data. In order to balance performance and reproducibility, we choose GLM3 - 32B. After conducting a series of extensive tests on multiple large language models, GLM3 - 32B emerged as the most suitable option for generating CEFI corpora. Unlike some other models with extremely large parameter sizes, GLM3 - 32B has a relatively moderate parameter scale. This characteristic makes it demand less computational power during the data generation process, which is highly beneficial for ensuring the efficiency and feasibility of our research. At the same time, the quality of the corpora generated by GLM3 - 32B meets the requirements of this task precisely. It can effectively generate data that closely resembles the characteristics of the original dataset in terms of semantic similarity, syntactic structure similarity, thematic consistency, coherence and fluency, and text style. This ensures that the additional data can be seamlessly integrated into the training process, enhancing the performance of the model without introducing excessive noise or biases.

Data validation: We prove the effectiveness of our generated data, which shares similar features with the original corpus. We choose popular methods or interfaces for verification and propose an effective method based on the characteristics of LLM and the definition of CEFI, which achieves good results.

Event Factuality Identification: We propose AMR-GCC, which divides CEFI into two sub tasks: document-level event classification and event collaboration. The large long text is encoded separately in the first sub task, analyzed using Abstract Meaning Representation (AMR) [6] to obtain important information, combining with a graph model to complete the classification task. When combining document-level events, we only use the titles and the classification results obtained and complete the task through mamba [4].

After completing the above submodules, we use the softmax function to obtain the final function value and evaluate its effectiveness.

3.3 Data Augmentation

Among the five categories, CT+ label which means the event will occur absolutely accounts for the majority, which is also consistent with the facts. After that, CT-, PS+, PS-, and Uu follow. In reality, this is also true for most news data. For Uu, which means ambiguity and unknown, this classification has limited significance and its frequency of appearance in news is too low, therefore its value is not significant.

Therefore, after balancing the quantity and meaning, we choose to generate data for the data with these labels: CT-, PS+, PS-. Meanwhile, for the sake of data validity, we cannot choose LLM with poor performance for fine-tuning. For the sake of reproducibility in the experiment, we cannot choose to use expensive or too large LLM. As a result, we chose GLM3-32B as the tool for this mission.

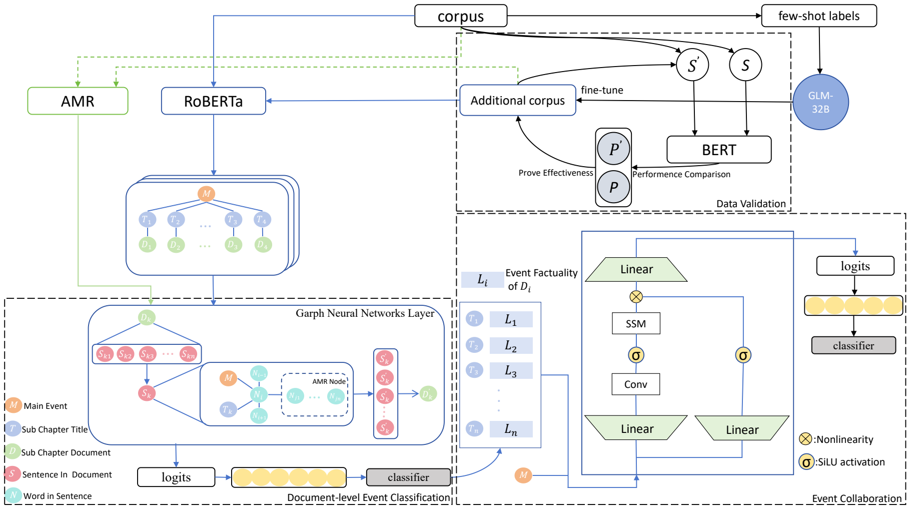

Fig. 2. An overall structural process of our method AMR-GCC for Cross-Documen Event Factuality Identification.

The method we use is LORA [5], which allows for less resource usage and shorter training time without compromising the fine-tuning results. We denote LLM as $P_{\Phi}(y|x)$ and its parameter as Φ. In the task, we set the prompt written in natural language as x and the answer of LLM as y, then a training dataset of context-target pairs can be written as $Z = (x_i, y_i)_{i=1,\ldots,N}$. In fine-tuning task, the initial weights of the model are set to Φ_0 and they are updated to $\Phi_0 + \triangle\Phi$ by repeatedly following the gradient to maximize the conditional language modeling objective:

$$\max_{\Phi} = \sum_{(x,y)\in Z} \sum_{t=1}^{|y|} \log(P_{\Phi}(y_t|x, y_{<t})) \tag{1}$$

The task-specific parameter increment Φ is set as $\Phi_0 + \triangle\Phi(\Theta)$, where $|\Theta| << |\Phi_0|$. Therefore, we obtain a smaller-sized set of parameters Θ that requires fewer resources and takes less time compared to full fine-tuning. The task of finding $\triangle\Phi$ thus becomes optimizing over Θ:

$$\max_{\Phi} = \sum_{(x,y)\in Z} \sum_{t=1}^{|y|} \log(P_{\Phi_0+\triangle\Phi(\Theta)}(y_t|x, y_{<t})) \tag{2}$$

Through this concept, we focus on the pre-trained weight matrix $W_0 \in R^{d\times k}$. We set the update of W_0 as $W_0 + \triangle W = W_0 + BA$, where $B \in R^{d\times r}$, $B \in R^{r\times k}$ and $r << min(d, k)$. A and B contain trainable parameters while W_0 does not receive gradient updates. For $h = W_0 x$, our modified forward pass yields:

$$h = W_0 x + \triangle x = W_0 x + BAx \tag{3}$$

We use a random Gaussian initialization for A and zero for B. The optimization method chosen is Adam.

After fine-tuning the large model for each individual label, we enable the model to generate data that matches the characteristics of our corpus. After performing these operations on each label we selected, we complete the experiment of data generation.

3.4 Data Validation

In CEFI, the purpose of constructing additional data is to solve the sparsity of the data and the uneven distribution with different labels limit. Therefore, in terms of syntactic structure similarity and thematic consistency, we will increase our attention. We need to ensure that the generated data enables our method to obtain features similar to the original data, rather than being used as noise to interfere with our experiments. So, we directly input the datasets before and after data augmention into BERT for comparison to judge the results.

We take the original training set as S and the generated data as S_{ai}, then the mixed training set is $S' = S \cup S_{ai}$. For the same test set T, when the style or text of S differs significantly from S_{ai}, S_{ai} will interfere with the final result of the method as noise, resulting in a decrease in effectiveness. When we can obtain similar features from S and S_{ai}, we will improve the performance on the data of few-shot labels by obtaining more features, ultimately enhancing the effectiveness of the method.

The details of the dataset can be obtained in the Experimental Setup module, and we use the most straightforward method to test whether the generated data has similar features to the original corpus. We only use five classification method of BERT on the same test set to classify S and S_{ai} separately after obtaining the results and obtain the results. Given real distribution p and predicted distribution q, we adopt the cross entropy loss function $\mathcal{L}$. The final loss function is as follows:

$$\mathcal{L} = -\sum_{x \in S} p(x) \ln q(x) \tag{4}$$

This step is employed to confirm the accuracy of the generated data. For data categorized under CT-, PS+, and PS-, their performance metrics show a notable improvement. Simultaneously, both macro and micro performance indicators also exhibit a significant increase.

Undoubtedly, the performance of the data we generated on few-shot data performs well, which verifies the effectiveness of our data and is the foundation of our method after implementation.

3.5 Event Factuality Identification

As illustrated by Fig. 2, we divide CEFI into two sub tasks based on this characteristic: document-level event classification and event collaboration, aiming to obtain the event factuality of document-level events and obtain the cross-document event factuality by combining document-level events. Therefore, we conduct AMR [6] analysis on the document-level event and construct a graph with the cross-document event to obtain the event factuality. Then we use mamba to jointly process the obtained information and obtain the cross-document event factuality.

Document-level Event Classification We record the cross-document event as M, the document-level event title set as $T = \{T_1, T_2, ..., T_n\}$, the corresponding document-level event text set $D = \{D_1, D_2, ..., D_n\}$, and the sentence set S corresponding to D. Record G as the global graph information obtained through AMR:

$$G = \{G_M, G_T, G_D\} = AMR(\{M, T, D\}) \tag{5}$$

In the formula, G_D includes the corresponding G_S. Afterwards, we obtain most of the nodes in the text and the information between them, which is of great help to us in obtaining the factual events of the document-level events.

For ease of representation, we introduce a symbol $[x|y]$, which is considered as concatenating x and y. We introduce a as a single-layer feedforward neural network and W as the weight matrix, and record that the word node in the sentence node is N. In the corresponding AMR structure, N_i represents the set of nodes adjacent or related to i. Therefore, the importance of node j to node i can be obtainable. For normalization processing, we use LeakyReLU as the activation function, which has fewer parameters and can increase the model's expressive power. Attention mechanism expression:

$$\alpha_{ij} = \frac{\exp(LeakyReLU(a^T([Wh_i|Wh_j])))}{\sum_{k \in Ni} \exp(a^T([Wh_i|Wh_k])))} \tag{6}$$

After getting the attention weights of each node in N_i for the current node i, we can get a new representation, seen as aggregating the information of nodes in N_i to obtain the updated representation of the current node:

$$h'_i = \sigma(\sum_{j \in Ni} \alpha_{ij} W h_j), j \in N_i \tag{7}$$

We take h_i as the representation of node i and use RoBERTa [8] for classification, calculate it using entropy loss function, obtain $\mathcal{L}_1$, and use it for training. Although BERT is the most widely-used baseline model, after testing, RoBERTa, as its modified version, is more suitable for our task. Therefore, we choose it as our encoding model.

For individual documents, we get their event factuality understanding of the cross-document events. Combined with titles, we obtain a set representation of T_i, D_i, where D_i represents the label of the cross-document event factuality in document-level event i. With the cross-document event, we proceed to the second sub task.

Event Collaboration As illustrated by Fig. 2, we record $x = T, D, M$ as input into mamba. The reason for choosing it is that mamba [4] has selective processing of input information and a simpler architecture, and its hardware-aware algorithms also reduce computing power requirements.

After passing through the linear layer, convolution operation, and SiLU activation, x enters the important SSM layer, which is the core of the second step.

First,we get $g_t = \sigma(Linera(x_t))$. For Selective State Space Model (SSM), the recursive equation is:

$$h_t = (1 - g_t)h_{t-1} + g_t x_t \tag{8}$$

In SSM, the selection mechanism reintroduces input-dependent dynamics, necessitating a hardware-aware algorithm that meticulously manages the materialization of expanded states within the more efficient tiers of the GPU memory hierarchy. $\varDelta, A, B, C$ are four trainable continuous parameters, and a major innovation here is to obtain discrete parameters $\overline{A}, \overline{B}$ through A, B:

$$\overline{A} = \exp(\varDelta A) \tag{9}$$

$$\overline{B} = (\varDelta A)^{-1}(\exp(\varDelta A) - I) \cdot \varDelta B \tag{10}$$

Discretization is profoundly intertwined with continuous-time systems, a relationship that can imbue these systems with supplementary attributes like resolution invariance [10] and the automatic guarantee of proper model normalization [11].

Afterwards, for 1-dimensional function or sequence $x_t \rightarrow y_t$ through an implicit latent state h_t, we update A, B to $\overline{A}, \overline{B}$, and through $h_t = \overline{A}h_{t-1} + \overline{B}x_t$, $y_t = Ch_t$, we get:

$$\overline{K} = (C\overline{B}, C\overline{AB}, ..., C\overline{A}^k\overline{B}, ...) \tag{11}$$

We obtain $y = x * \overline{K}$, the output representation of x in 1-dimensional, and for the real result, we obtain it through a higher dimensional latent state in the same way. Combined with it, we obtain the probability function q.

We transform the initial factual annotation of the event into the hot format to derive the original true probability function p, representing the distribution of predicted values. All samples are collectively represented by S, while the individual sample employed is denoted as x. Given that this task pertains to classification, we adopt the cross entropy loss function for our calculations. The loss $\mathcal{L}_2$ obtained, we train the model and complete the experiment.

4 Experimentation

4.1 Dataset construction

We briefly introduce the data sources, events selection, annotation rules, and specific information of our dataset.

Data Sources: We crawled relevant news reports on the same theme from ChinaDaily and annotated them manually. The news reports primarily originate from reputable official news websites, including CCTV News, Huanqiu, and Sina, thereby guaranteeing the information's reliability.

Events Selection: Cross-document events are events in the theme that have been stripped of redundant descriptions, leaving only subject, verb, and object. And events must be articulated in affirmative sentence structures, devoid of any terms that imply factual biases, such as negative or speculative indicators.

Annotation Rules: The annotation tool we use was doccano, an open-source text annotation tool commonly used in NLP task annotation. The specific format of one sample includes cross-document event and several corresponding texts under the document-level event. The title of a document-level event is the title of a news report, and the text is

Table 1. Dataset Statistics

	Total	CT+	CT-	PS+	PS-	Uu
number	4294	3701	308	230	43	12
percent	100	86.1	7.2	5.3	1.1	0.3
additional number	2946	0	1183	1176	587	0
final number	7240	3701	1491	1406	630	12
final percent	100	51.0	20.1	19.2	8.5	0.2

all the documents crawled up. Each event factuality was annotated repeatedly by two people. If there was a conflict, a third party was added for annotation. If all three people had different annotations, we discussed them together and provided judgements.

Specific Information: Table 1 shows the dataset we created, a newer English corpus for CEFI, which consists of 4294 samples. For this dataset, as it is manually annotated, the majority of events with a label of CT+ account for the total number, which is consistent with the fact. It is precisely to solve this problem that we expand the data. Table 1 also lists the effects after completing data augmentation. It can be seen that although events labeled as CT+ still dominate, the proportion of the other three important label events significantly increases.

4.2 Baselines

We use the following methods as baselines:

- Bert [2]: The most widely used and direct method currently, designed to pre-train deep bidirectional representations from the unlabeled text by joint conditioning on both left and right context in all layers.
- ULGN [1]: A method combining a Local Uncertainty Estimation module to model the uncertainty of local information and an Uncertain Information Aggregation module to leverage the global structure for integrating the local information, treats multiple documents as a single document.
- TARA [14]: A method proposing Tailored AMR Graph, which compresses less informative subgraphs, integrates span information, highlights surrounding events, and employs graph neural networks for link prediction to identify event arguments.
- GLM-130B [15]: A powerful large language model, as a bilingual pre-trained language model with 130 billion parameters, obtaining results by fine-tuning. The input is multiple documents in stages when processing a single sample.

4.3 Overall Results

Table 2 shows the specific results. Due to the scarcity and insignificant significance of Uu, we mainly examine the performance of CT+, CT-, PS+, and PS-. We can draw the following conclusions:

Our method excels on this corpus, achieving notable improvements across all six performance indicators. Particularly noteworthy is the substantial enhancement in Macro-F1, which highlights our method's superior handling of class imbalances. Additionally,

Table 2. Experimental results, we use F1-Score to evaluate the performance of the five categories CT+, CT-, PS+, PS-, Macro–F1, Micro-F1.

Methods	CT+	CT-	PS+	PS-	Macro-F1	Micro-F1
BERT	92.36	38.09	35.29	0	33.15	85.78
ULGN	93.26	42.86	42.31	0	35.69	87.15
TARA	94.30	46.51	46.81	18.18	41.16	88.11
GLM-130B	94.43	47.83	45.83	25.0	42.62	88.34
Ours	**95.89**	**56.52**	**52.17**	**36.36**	**48.59**	**90.49**

a significant boost in Micro-F1 further underscores the method's enhanced precision and recall. Together, these improvements strongly validate the effectiveness and robustness of our approach.

F1 Score: The discrepancy between Macro-F1 and Micro-F1 values across all methods is notably substantial, a phenomenon we primarily attribute to the inherent characteristics of the dataset, as influenced by the specific task definition of CEFI. The data's inherent sparsity, coupled with the uneven distribution of labels and varying label constraints, presents inevitable challenges that significantly impact these performance metrics.

CT+: In CT+, our improvement is relatively modest, primarily because our task is specifically designed to enhance the performance of data with fewer labels. We did not generate any additional data for CT+, and given that its initial F1 value was already quite high, the scope for further improvement was inherently limited, resulting in a smaller relative enhancement.

CT- And PS+: Our improvement in CT- and PS+ with low quantity labels is undeniably substantial. We attribute this significant advancement to the high reliability of our generated data, which empowers our method to extract richer semantic features. This capability represents a major breakthrough, underscoring the effectiveness of our approach in enhancing performance even in data-scarce scenarios

PS-: In PS-, all methods exhibited subpar performance, a situation we primarily attribute to corpus-related issues such as an insufficient sample size and poor data quality. For our method, the limited availability of data significantly hampered the fine-tuning process and the enhancement of method robustness, Consequently, this constraint led to only a modest improvement in performance.

Scalability: Since the largest model deployed in our method is the fine-tuned GLM-32B, the required hardware requirements and computing power consumption are not that large. After achieving improved performance enhancements, this is obviously acceptable. Meanwhile, when the dataset is enlarged, during the stage of generating the dataset, the dataset can be screened to reduce its scale before fine-tuning. Moreover, we have split the complex CEFI into two steps, ensuring that the method will not produce errors due to the increase in the size of the dataset.

Comprehensive Performance: Overall, each component of our method demonstrates significant effectiveness. The augmentation of low-label corpus data plays a crucial role in substantially boosting the macro-F1 metric, highlighting improved performance in handling class imbalances. Furthermore, our approach to capturing the unique

Table 3. Ablation study result, the arrow ↓ represents the decrease.

Methods	CT+	CT-	PS+	PS-	Macro-F1	Micro-F1
w/o at	↓0.53	↓10.78	↓14.45	↓18.56	↓11.65	↓1.36
w/o ft	↓0.42	↓9.83	↓12.46	↓18.56	↓10.62	↓1.04
w/o am	↓1.34	↓4.76	↓7.34	↓5.78	↓4.85	↓1.02
w/o gl	↓0.85	↓5.69	↓6.28	↓7.86	↓4.14	↓0.87
w/o ma	↓0.72	↓6.39	↓7.12	↓5.45	↓3.93	↓0.64

structural elements of documents and consolidating key primary events proves exceptionally beneficial, leading to notable enhancements in micro-F1 scores. This comprehensive strategy underscores the method's ability to address both broad and specific aspects of text processing, resulting in overall performance gains.

4.4 Ablation Study

We conduct the following ablation studies to validate the effectiveness of our proposed components: 1) w/o at: removing all additional AI text added to the training set. 2) w/o ft: removing the section of fine-tuning LLM and allowing it to directly generate AI text. 3) w/o am: removing AMR and replacing it as usual adjacent node relationship. 4) w/o le: removing Graph Neural Networks Layer and directly encoding for logits calculation. 5) w/o ma: removing mamba which joins the classification results of the sub task and directly using `concat` for concatenation. The ablation study results are shown in Table 3.

Based on the results, it is evident that the removal of any component results in a decline in performance. Removing "at" or "ft" will cause the model to be unable to obtain sufficient semantic features for few-shot labels, resulting in a significant decrease in Macro-F1. Removing "am" or "gl" will make the model insensitive to the input data features, making it difficult to obtain its correct features, resulting in a decrease in model performance. Removing "ma" will make the processing of long texts more difficult, making it more difficult to capture key information in long texts. Therefore, ablation Study proves that every step of our method is indispensable.

4.5 Case Study

In this section, we examine specific examples to explore how the two specific steps of AMR-GCC work.

As shown in Fig. 3, due to length limitations, we list the titles and sentences that indicate the event factuality of events in each document in the example. The central issue of the event revolves around whether the town will undergo a name change to "Kush". To ascertain the event factuality of each document, we leverage colored words, specifically the expressed attitudes of individuals. Our approach involves conducting Document-level Event Classification, wherein we construct Abstract Meaning Representation (AMR) maps within each document. These AMR maps are then integrated with the broader context of the cross-document event.

By meticulously analyzing the colored important words, we derive the event factuality identification for all individual documents. Subsequently, we engage in event

Cross-document Event: This Colorado town change its name to ‘Kush’.	
Cross-Document Event Factuality: PS+	
[D1:PS+]	The town is now considering changing its name to ‘Kush’. The cannabis industry has had a major impact on the tiny town of Moffat, Colorado – so much that the town is now considering changing its name to ‘Kush’.
[D2:PS+]	Biggio suggested the name change during the public comment period at a meeting. “It felt like almost the whole town came out to share their side,” Foxx said.
[D3:PS-]	Foxx says that not everyone was enthusiastic about the possible name change. Some residents wanted to keep the town’s “history, heritage, and identity” embodied in its name.
[D4:PS+]	The next step in transforming Moffat to Kush would be for a resident to start a petition. Foxx thinks the possibility of changing the town’s name is “exciting”.

Fig. 3. An example for Case Study

collaboration, merging the findings from three documents classified as PS+ and one classified as PS-, along with their respective titles. This comprehensive integration allows us to determine the cross-document event factuality across the entire corpus of texts.

Ultimately, after thorough analysis and synthesis, the resultant cross-document event factuality is classified as PS+. This conclusion is derived from the collective evidence and the nuanced interplay of attitudes and expressions present in the documents, highlighting the robustness of our method in handling complex, multi-document event factuality assessments.

4.6 Error Analysis

In this section, we select a representative error case and perform a detailed analysis.

As shown in Fig. 4, although numerous texts demonstrate that cross-document event factuality is PS-, the incorrect judgment is largely due to the initial biased summary favoring PS+ and the weak negative connotations in the text. This initial bias can overshadow the subtle negative indicators, leading to a misinterpretation of the overall evidence. In D1, despite the internal conflicts within the text, the overarching bias of the narrative still tilts towards the idea that a rising dollar is beneficial for the Southeast Asian economy. In D2, D3 and D4, the text indicates the uncertainty of the event from different perspectives, but because it does not directly point it out but indirectly explains it from multiple angles, our method cannot obtain its true intention.

Overall, in the conflict between PS- and PS+, our method fails to accurately distinguish which one is correct. This inadequacy arises primarily from the direct and explicit expression of PS+, which tends to overshadow the more subtle and less pronounced expression of PS-. The stark contrast in the clarity of these expressions complicates the task of making a definitive judgment, thereby highlighting a limitation in our method’s ability to handle nuanced distinctions between the two.

Cross-document Event: A rising dollar is good for Southeast Asia.	
Error Label: PS+	**True Label**: PS-
[D1:PS+]	Strong dollar cuts both ways in SE Asia. A surging US dollar may benefit Southeast Asia's export and tourism sectors, but such gains could be undercut by costlier imports.
[D2:PS-]	But Southeast Asian countries are also dependent on imported fuel and food and their weaker currencies have led to higher consumer prices. "Weaker local currencies tend to favor export-oriented countries that make their exports cheaper in international markets and increase the local currency equivalent of their export revenues in US dollar," said Michael Ricafort, chief economist at Rizal Commercial Banking Corp in Manila.
[D3:PS-]	However, for net importing countries, weaker local currencies will increase import costs and boost inflation, Ricafort said. But even net exporters cannot keep inflation at bay. Such is the dilemma in Malaysia, Southeast Asia's third-biggest economy.
[D4:PS-]	Inflation in Southeast Asia's largest economy rose to a seven-year high of 5.95 percent in September. But these export revenues cannot offset the impact of higher food prices.

Fig. 4. An example for Error Analysis

5 Conclusion

In this paper, we propose a new concept, Cross-Document Event Factuality Identification (CEFI). We manually annotate a dataset, which faces the sparsity of the data and the uneven distribution with different labels limit, and fine-tune GLM-32B to solve the problem. On this basis, we conduct Data Validation to ensure the reliability of the generated corpus. AMR-GCC is proposed as a new method that divides tasks into two sub tasks. The use of AMR to parse semantic structures and combine graph models to obtain factual information about sub discourse events improves the robustness of long text tasks. Afterwards, mamba is used to combine all document-level events with the cross-document event, reducing computing power requirements and improving effectiveness. The experimental results demonstrate our method achieves state-of-the-art performance.

Acknowledgment

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62276177, 62006167, 62376181, 62176175), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 24KJB520036), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

1. Cao, P., Chen, Y., Yang, Y., Liu, K., Zhao, J.: Uncertain local-to-global networks for document-level event factuality identification. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021. pp. 2636–2645 (2021), https://doi.org/10.18653/v1/2021.emnlp-main.207
2. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. pp. 4171–4186 (2019), https://doi.org/10.18653/v1/n19-1423
3. Gao, L., Liu, Y., Zhu, J., Yu, Z.: A cognitively inspired multi-granularity model incorporating label information for complex long text classification. Cogn. Comput. **16**(2), 740–755 (2024), https://doi.org/10.1007/s12559-023-10237-1
4. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. CoRR **abs/2312.00752** (2023), https://doi.org/10.48550/arXiv.2312.00752
5. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations, ICLR 2022 (2022), https://openreview.net/forum?id=nZeVKeeFYf9
6. Kouris, P., Alexandridis, G., Stafylopatis, A.: Text summarization based on semantic graphs: an abstract meaning representation graph-to-text deep learning approach. J. Big Data **11**(1), 95 (2024), https://doi.org/10.1186/s40537-024-00950-5
7. Langhe, L.D., Clercq, O.D., Hoste, V.: Enhancing unrestricted cross-document event coreference with graph reconstruction networks. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024. pp. 6122–6133 (2024), https://aclanthology.org/2024.lrec-main.541
8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
9. Murayama, T.: Dataset of fake news detection and fact verification: A survey. CoRR **abs/2111.03299** (2021), https://arxiv.org/abs/2111.03299
10. Nguyen, E., Poli, M., Faizi, M., Thomas, A.W., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C.M., Bengio, Y., Ermon, S., Ré, C., Baccus, S.: Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. In: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023 (2023), http://papers.nips.cc/paper_files/paper/2023/hash/86ab6927ee4ae9bde4247793c46797c7-Abstract-Conference.html
11. Orvieto, A., Smith, S.L., Gu, A., Fernando, A., Gülçehre, Ç., Pascanu, R., De, S.: Resurrecting recurrent neural networks for long sequences. In: International Conference on Machine Learning, ICML 2023. Proceedings of Machine Learning Research, vol. 202, pp. 26670–26698 (2023), https://proceedings.mlr.press/v202/orvieto23a.html
12. Qian, Z., Li, P., Zhou, G., Zhu, Q.: Event factuality identification via hybrid neural networks. In: Neural Information Processing - 25th International Conference, ICONIP 2018. Lecture Notes in Computer Science, vol. 11305, pp. 335–347 (2018), https://doi.org/10.1007/978-3-030-04221-9_30
13. Qian, Z., Li, P., Zhu, Q., Zhou, G.: Document-level event factuality identification via adversarial neural network. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. pp. 2799–2809 (2019), https://doi.org/10.18653/v1/n19-1287

14. Yang, Y., Guo, Q., Hu, X., Zhang, Y., Qiu, X., Zhang, Z.: An amr-based link prediction approach for document-level event argument extraction. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023. pp. 12876–12889 (2023), https://doi.org/10.18653/v1/2023.acl-long.720
15. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., Tam, W.L., Ma, Z., Xue, Y., Zhai, J., Chen, W., Liu, Z., Zhang, P., Dong, Y., Tang, J.: GLM-130B: an open bilingual pre-trained model. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023 (2023), https://openreview.net/forum?id=-Aw0rrrPUF
16. Zhang, H., Qian, Z., Li, P., Zhu, X.: Evidence-based document-level event factuality identification. In: PRICAI 2022: Trends in Artificial Intelligence - 19th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2022. Lecture Notes in Computer Science, vol. 13630, pp. 240–254 (2022), https://doi.org/10.1007/978-3-031-20865-2_18
17. Zhang, H., Qian, Z., Zhu, X., Li, P.: Document-level event factuality identification using negation and speculation scope. In: Neural Information Processing - 28th International Conference, ICONIP 2021. Lecture Notes in Computer Science, vol. 13108, pp. 414–425 (2021), https://doi.org/10.1007/978-3-030-92185-9_34
18. Zhang, Z., Elfardy, H., Dreyer, M., Small, K., Ji, H., Bansal, M.: Enhancing multi-document summarization with cross-document graph-based information extraction. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023. pp. 1688–1699 (2023), https://doi.org/10.18653/v1/2023.eacl-main.124