Developing A Novel Bidirectional Sparse Graph Attention Adaptor for Evidence-Based Fact-Checking

Anonymous ACL submission

Abstract

Evidence-based Fact-checking aims to verify or debunk a claim with evidence given and has benefited from Large-Language-Model (LLM) advancements in text understanding. However, autoregressive LLMs suffer from their unidirectional nature, known as "Reversal Curse", causing their performance to be unsatisfactory. Therefore, in this paper, we propose to utilize bidirectional attention as an external adapter for two-way information aggregation. Further, we leverage hierarchical sparse graphs to reduce the noise impact of attention and an efficient feature-compression mechanism to reduce the number of adaptor parameters. Experimental results on both English and Chinese datasets demonstrate the significant improvements achieved by our proposed approach and its state-of-the-art performance in the Evidence-based Fact-checking task. The code will be available on GitHub.

1 Introduction

In the face of the growing spread of misleading information in the real world, fact-checking becomes necessary to turn the tides of misinformation (Vosoughi et al., 2018; Khan et al., 2021). Evidence-based Fact-checking (EBFC) seeks to verify or debunk a claim with evidence given, benefiting from the development of Large Language Models (LLMs), such as GPT and Llama (Cao et al., 2023; Quelle and Bovet, 2023; Cheung and Lam, 2023).

However, LLMs struggle to judge the claim after learning the evidence that swaps the order, known as the "Reversal Curse" (Grosse et al., 2023; Berglund et al., 2023), due to the unidirectional nature of the autoregressive LLMs. As an example presented in Table 1, with the order of "boiling water" and "dishes" in evidence swapped compared to the claim, GPT-4 made a wrong prediction. Our preliminary analysis of the Evidence-based Fact-checking dataset CHEF (Hu et al., 2022) showcases that 48.31% of inaccuracies in the outcomes produced by GPT-4 can be attributed to the Reversal Curse.

Verify or debunk the claim with the evidence given. **The Claim**: Dishes cannot be sterilized with boiling water. **Evidence**: ... Evidence 4: Thus, boiling water cannot sterilize the dishes. ...
Dataset: CHEF; **ID**: 686; **Label**: Supported.
GPT-4 Prediction: Refuted. **GPT-4 Response**: ... Evidence 4 is a statement that contradicts the claim, stating that boiling water cannot sterilize the dishes. ...

Table 1: A Reversal Curse example of the Evidence-based Fact-checking task, where the statement in the claim is reversed to the selected statement in evidence.

Various attempts have been made to modify training setups (e.g., scaling model and data size) for LLMs to alleviate the Reversal Curse but failed to exhibit significant performance promotion (Grosse et al., 2023; Berglund et al., 2023). As LLMs may store facts differently depending on their direction (Meng et al., 2023), the "Reversal Curse" is an inborn defect of autoregressive models. In such a case, we explore designing a bidirectional adapter to overcome this drawback. Inspired by the human fact-checker gathers related evidence back and forth to understand the sentence meaning, we build new bidirectional attention in the linear layers of the Transformer self-attention module (Vaswani et al., 2023).

The adapter has been proposed to adapt LLMs for multiple downstream applications like reasoning (Houlsby et al., 2019), where adapters freeze the original model and add a few additional parameters for fine-tuning. Previous research (Hu et al., 2021) demonstrates that adapters achieve the best results when adapting to the Query and Value matrices of self-attention. Nevertheless, introducing bidirectional attention in Query may break the autoregressive Query-Key mask of LLMs. Following

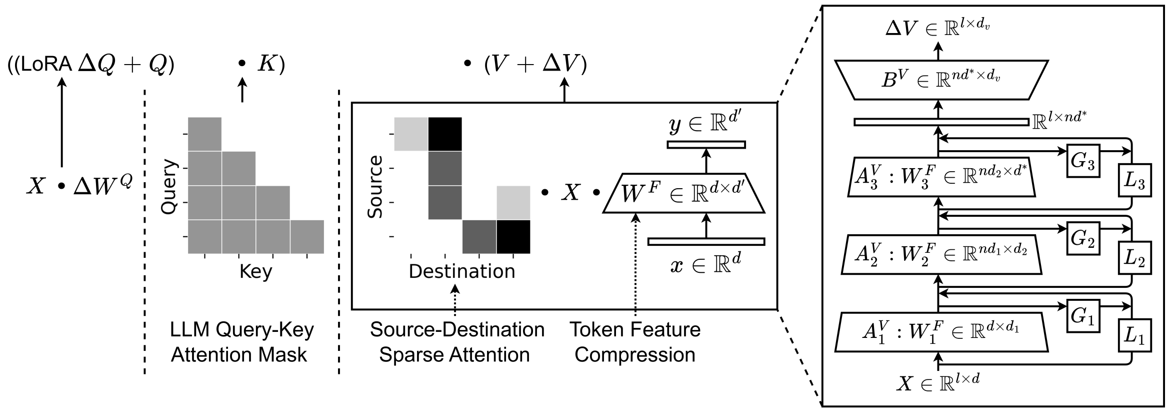

Figure 1: The framework of our proposed bidirectional sparse graph attention adaptor.

these two conclusions, our framework adapts Value to build bidirectional attention as shown in Figure 1. Furthermore, our method adapts Query with LoRA (Hu et al., 2021) to refresh Query-Key pairs for fine-tuning.

Our adaption models new bidirectional attention on sparse graphs, taking tokens as nodes and building attention with directed edges. Sparse means each token only pays attention to a few tokens with the most relevant information, which is critical to understanding the text (Zhao et al., 2019). We design three sparse graphs with different receptive fields and leverage a hierarchical structure with smaller receptive field graphs as input to larger graphs, aiming to merge local and global information in each layer. At the same time, skip connections and gate units are designed to balance the ratio of our bidirectional information injection to capture local and global dependencies (Cho et al., 2014).

In addition, our approach reduces the adapter parameters through a feature-compression mechanism on token representations for efficient adaption and further sparse feature selection. The feature dimension will be reduced gradually through each layer in the hierarchical structure, and finally, our framework splices a feature-decompression matrix for output.

In summary, in this work, we develop the novel **B**idirectional **S**parse **G**raph **A**ttention **A**daptor for evidence-based fact-checking (**BSGAA**). Our approach achieves state-of-the-art (SOTA) performance on both English and Chinese datasets. The main contributions include:

- We propose a bidirectional attention adapter to model two-way relations, representing the pioneering attempt to combine bidirectional information modeling with autoregressive LLMs.

- We develop a hierarchical sparse graph structure and feature-compression mechanism to make the adaption robust and efficient.

- Experimental results demonstrate that our method achieves SOTA performance, outperforming the GPT-4 on the Evidence-based Fact-checking task.

2 Methodology

2.1 Task Description and Overview

Evidence-based fact-checking (EBFC) (Augenstein et al., 2019) aims to verify or debunk the factual veracity of the questioned claim with several pieces of evidence retrieved by the automatic ranker or human annotator. The output will be three possible labels: SUP (Support), REF (refute), or NEI (not enough information).

For the convenience of notation, we use X, Q, K, and V to denote the Input, Query, Key, and Value and W^Q, W^K, and W^V as corresponding projection matrices in the LLM self-attention modules. As shown in Figure 1, we develop bidirectional sparse graph attention adaption of V to model two-way relations for information aggregation and utilize LoRA Q adaption to refresh Query-Key pairs for fine-tuning. The adapted attention mechanism (Vaswani et al., 2023) can be represented as:

$$\mathrm{Attn}(X, W^Q, W^K, W^V) \tag{2.1}$$
$$= \mathrm{softmax}\left(\frac{(Q + \Delta Q)K^{\mathsf{T}}}{\sqrt{d_k}}\right)(V + \Delta V).$$

d_k and d_v are the Key and Value dimensions of the LLM. Our model employs (1) Bidirectional Attention to model two-way relations and Sparse Graph to improve the concentration of attention in Section 2.2 and (2) Hierarchical Structure to merge local and global information in each layer and Feature-Compression Mechanism to reduce adapter parameters in Section 2.3.

2.2 Bidirectional Sparse Graph Attention

In this section, we propose to build new bidirectional attention, and we want the attention to be sparse for less noise impact. We leverage sparse graphs to better model sparse attention, taking tokens as nodes and building attention with directed edges (Velickovic et al., 2017). In this way, the attention of the i-th token is calculated only with its first-order neighbor (Sedgewick and Wayne, 2011) tokens $j \in \mathcal{N}_i$.

To distinguish attention symbols in the adapter from those in the LLM, we use Source (S) as Query, Destination (D) as Key, and Feature (F) as Value in the adapter. Following Vaswani et al. (2023), our adaption utilizes a multi-head attention mechanism, and n is the number of attention heads. To elaborate our approach, we demonstrate the m-th layer of the three-layer hierarchical structure for general description, and each layer takes the output of its former layer as input.

Denote the input H_{m-1} of the m-th layer as:

$$H_{m-1} \in \mathbb{R}^{l \times d_{m-1}}, m = 1, 2, 3, H_0 = X, d_0 = d.$$

l is the token number of the input text and d_{m-1} is the feature dimension of the input. X is the input of the LLM self-attention module and d is the feature dimension of X.

We start with building the Query S, Key D, and Value F attention matrices. Our approach first builds the Value F, utilizing the projection matrix $W_m^F \in \mathbb{R}^{d_{m-1} \times d_m}$.

$$F_m = [F_{m1}, \cdots, F_{ml}] = X W_m^F. \tag{2.2}$$

d_m is the output feature dimension of each attention head. d_m can be freely altered for compression or decompression, and we will discuss this in Section 2.3. With Value F as input, we calculate Query S and Key D with projection matrix W_m^S and W_m^D.

$$S_m = \tanh(F_m) W_m^S, W_m^S \in \mathbb{R}^{d_m \times 1}, \tag{2.3}$$

$$D_m = \tanh(F_m) W_m^D, W_m^D \in \mathbb{R}^{d_m \times 1}, \tag{2.4}$$

We leverage the nonlinear activation function $\tanh$ to prevent S, D, and F from forming linear relationships with each other, therefore better leveraging and capturing the graph structure information (Qiu et al., 2018).

Our approach initializes the sparse graph with a receptive field r_m constraint.

$$j \in \mathcal{N}_i \iff |i - j| \leq r_m. \tag{2.5}$$

Now we calculate the attention $E_m \in \mathbb{R}^{l \times l}$ of the directed edges $i \to j$ on the graph.

$$e_{mij} = \underset{\alpha=0.2}{\text{LeakyReLU}}(S_{mi} + D_{mj}), \tag{2.6}$$

$$E_{mij} = \underset{j \in \mathcal{N}_i}{\text{softmax}}(e_{mij}). \tag{2.7}$$

Our framework calculates the attention score e_{mij} by adding Query S_{mi} of the i-th token and Key D_{mj} of the j-th token and then normalizes e_{mij} with softmax. Our approach adds Query and Key other than point-wise multiplication, such that the magnitude of S and D does not affect the gradient descent of each other. According to our experimental results, the summation enhances the concentration of attention through implicit selection during training, and the gradient descent speed can still be maintained under sparse situations.

Finally, we use the ELU output activation function to obtain the output $\hat{H}_m$ with the following expressions.

$$\hat{H}_m = \text{Concat}(\text{ELU}(\sum_j E_{mij} F_{mj})). \tag{2.8}$$

In summary, our bidirectional sparse attention fuses the information of token $j \in \mathcal{N}_i$ into token i.

2.3 Hierarchical Structure and Feature-Compression Mechanism

In this section, we design three sparse graphs with different receptive fields, stacking them in a hierarchical structure with a pass-through and a feature-compression mechanism.

We construct a hierarchical sparse graph stack to combine local and global information in each layer, where the representations of the lower layer serve as the input to the higher layer. This stack applies three granularities of receptive fields for three layers in Inequation (2.5), where lower layers concentrate on a narrow range around each token to get relatively local information and higher layers focus on a broader range.

$$|i - j| \leq r_m, r_1 < r_2 < r_3.$$

This way, bidirectional relations between tokens caring for different ranges are modeled.

In addition, our framework employs a pass-through mechanism with linear layers $L_m \in \mathbb{R}^{d_{m-1}\times d_m}$, utilizing a gate control mechanism with linear gates $G_m \in \mathbb{R}^{d_m\times 1}$ to balance the ratio of our sparse bidirectional information injection.

$$\hat{H}_m = A_m^V(H_{m-1}), \tag{2.9}$$

$$\begin{aligned} H_m &= (1 - \mathrm{sigmoid}(\hat{H}_m G_m)) * H_{m-1}L_m \\ &\quad + \mathrm{sigmoid}(\hat{H}_m G_m) * H_m. \end{aligned} \tag{2.10}$$

$$m = 1, 2, 3, H_0 = X.$$

We use A_m^V to denote all calculations from Equation (2.2) to Equation (2.8) in each layer. The "$*$" is the broadcast multiplication in Equation (2.10).

Furthermore, our method reduces adaption parameters through a feature-compression mechanism to make adaption efficient. As stated in Section 2.2 Equation (2.2), we alter d_m for feature dimension compression on the hierarchical graphs. Each layer of our hierarchical adapter smoothly projects the input to a smaller subspace with Value projection $W_m^F \in \mathbb{R}^{d_{m-1}\times d_m}, m = 1, 2, 3$ in Equation (2.8), as shown in Figure 1, where $d^* = d_3 < d_2 < d_1 << min(d_0 = d, d_v)$. To align the dimensions of output H_3 and V, we splice a decompression matrix multiplier $B^V \in \mathbb{R}^{nd^*\times d_v}$.

$$\Delta V = H_3 B^V. \tag{2.11}$$

Meanwhile, this feature-compression mechanism forces clipping out the useless part of attention, thus making the attention more sparse and spontaneously learning the sparse information.

In summary, our proposed hierarchical structure merges local and global information and meticulously maintains the balance of bidirectional information injection. The feature-compression mechanism reduces the adapter parameters and makes the attention more sparse through feature selection.

2.4 Training and Answer Prediction

In this section, we define the loss of our model here and summarize our training and answer prediction approach. Our approach utilizes the feature z of the last token in the LLMs and uses a linear layer to project it into a 3-dimensional score vector $\hat{y}$.

$$\hat{y} = \mathrm{Score}(z) = zS, \tag{2.12}$$

where $S \in \mathbb{R}^{d\times 3}$. We then utilize the 3-dimensional score vector $\hat{y}$ to make our 3-way prediction for Evidence-based Fact-checking.

$$y^* = \mathrm{softmax}(\hat{y}), \tag{2.13}$$

where y^* denotes the predicted probability of categories.

Our framework freezes all the parameters of the LLMs and only updates the parameters of W^F, W^S, W^D, G_m, L_m, and B^V of feature-compression sparse graph attention layers and A^Q, B^Q of LoRA Q adaption. Our method leverages backpropagation with cross-entropy label loss $\mathcal{L}_{CE}$ for training.

$$\mathcal{L}_{CE} = \mathrm{CrossEntropy}(y^*, y), \tag{2.14}$$

where y is the true label.

For answer prediction, we consider the category with the largest probability in y^* as the predicted label of our model.

$$y_{pred} = \mathrm{argmax}(y^*), \tag{2.15}$$

where $y_{pred} \in \{0, 1, 2\}$ is the predicted answer of inference.

3 Experiments

3.1 Dataset

To investigate the effectiveness of the proposed method, we conducted our research on Evidence-based Fact-checking datasets FEVER (English) (Thorne et al., 2018) and CHEF (Chinese) (Hu et al., 2022). FEVER (Thorne et al., 2018) consists of 185,445 synthetic claims by altering sentences extracted from introductory sections of Wikipedia pages and combining several sentences to form the necessary evidence. CHEF (Hu et al., 2022) consists of 10,000 real-world claims collected from 6 Chinese fact-checking websites and uses several corresponding source documents retrieved through Google Search API as evidence. Both of their labels have three classes, which are supported (0 or SUP), refuted (1 or REF), and not enough information (2 or NEI).

The lengths of training sets of FEVER and CHEF are 145449 and 8002, respectively. For the comparison between the performance of our model in FEVER and CHEF, we randomly chose 8002 examples in FEVER to build the dataset that we used in our experiments in our paper. We take the top 5 pieces of evidence for each claim in both datasets.

Our framework leveraged the given golden evidence and randomly sampled sentences as evidence

Method	Model	Trainable Parameters	FEVER LA (%)	CHEF LA (%)	CHEF F1 (%)
X-Fact (Gupta and Srikumar, 2021)	mBERT-base (Devlin et al., 2019)	125M	-	63.48†	62.47†
GEAR (Zhou et al., 2019)	BERT-base (Devlin et al., 2019)	110M	71.60	-	-
KGAT (Liu et al., 2020)	BERT-base	110M	85.15*	64.37†	62.58†
TwoWingOS (Yin and Roth, 2018)	TwoWingOS	NA	75.99	67.46‡	64.31‡
CHEF (Hu et al., 2022)	BERT-base	110M	-	69.12	65.26
BEVERS (DeHaven and Scott, 2023)	RoBERTa-large (Liu et al., 2019)	355M	79.39	-	-
ProoFVer (Krishna et al., 2022)	BART-large (Lewis et al., 2020)	400M	79.47	-	-
ReRead (Hu et al., 2023b)	BERT-base	110M	-	70.87	68.78
Cao et al. (2023) (zero-shot)	GPT-3.5 (gpt-3.5-turbo)	0	-	35.14	33.51
Cao et al. (2023) (zero-shot)	Llama-2 (7B)	0	-	31.93	28.58
GPT-4 (zero-shot)	GPT-4 (gpt-4-1106-preview)	0	93.91*	68.69	64.17
LoRA (fine-tuned, ours)	Llama-2 (7B)	5M	94.29*	70.17	66.59
BSGAA (w/o feature-compression)	Llama-2 (7B)	150M	94.50*	71.37	68.61
BSGAA	Llama-2 (7B)	5M	**95.08***	**72.97**	**70.05**

Table 2: Evidence-based Fact-checking results on FEVER (English) and CHEF (Chinese). $*$ indicates the results produced with golden evidence on FEVER. † indicates the results reproduced on CHEF by Hu et al. (2022). ‡ indicates the results reproduced on CHEF using graph-based model KGAT (Liu et al., 2020) by Hu et al. (2022).

Label	CHEF Dataset train	CHEF Dataset dev	CHEF Dataset test
SUP	319(11.09%)	37(11.11%)	38(11.41%)
REF	783(18.00%)	57(17.12%)	57(17.12%)

Table 3: Statistics of instances with no golden evidence in CHEF.

of NEI claims for FEVER. As shown in Table 3, while CHEF has instances with no golden evidence to test intrinsic knowledge of models, we employ the automated retrieval evidence retrieved by Hybrid Ranker (Shaar et al., 2020; Hu et al., 2022) for CHEF. Our statistics also show that CHEF has 45 (13.51%) SUP instances and 60 (18.02%) REF instances with Reversal Curse for evidence retrieved, and we packaged these instances into a new dataset CHEF-RC (CHEF-Reversal Curse).

Following prior efforts (Thorne et al., 2018; Augenstein et al., 2019; Liu et al., 2020; Hu et al., 2022), we adopt label accuracy (LA) as FEVER evaluation metrics, and label accuracy (LA), macro F1 score of the label (shown as F1) as CHEF evaluation metrics to assess the performance of our model. We also apply label precision (P) and recall (R) for each classification category in the following analyses.

3.2 Experimental Settings

We adopted Llama-2-7B (Touvron et al., 2023) for our method, and our experiments were run on 1 NVIDIA RTX-3090 GPU. For simplicity, we conduct adaptions only on the 32nd layer. The feature dimension of Llama-2-7B is 4096, and the output dimension of each layer of our hierarchical bidirectional attention adapter is sequentially 256, 16, and 4. Our model is trained for a maximum of 5 epochs using the AdamW optimizer, which features an initial learning rate of 2e-4, a weight decay of 0.01, and a warm-up rate of 0.05. The batch size of our model is set to 8, and we use the dropout technique with a dropout rate of 0.1 for regularization.

Since LoRA is an efficient adaption framework, we set up a comparative LoRA baseline with the settings above, except the intermediate dimension is 10 to match the total parameters of BSGAA. To explore the Evidence-based Fact-checking ability of GPT, we conducted a preliminary attempt to utilize the zero-shot GPT-4 model to deal with the task. For experiments on GPT-4, We utilize gpt-4-1106-preview API and set every parameter by default to do preliminary research on its performance of Evidence-based Fact-checking.

3.3 Baselines

To show the effectiveness of our model, we compare our results with other baselines. Since many previous works use small models as classifiers, they are not competitive with LLMs, and we only list some of them as baselines.

X-Fact (Gupta and Srikumar, 2021) used an attention-based evidence aggregator (Attn-EA) to emulate the evidence aggregation behavior of human fact-checkers. GEAR (Zhou et al., 2019) proposed a graph-based evidence aggregation to transfer information on evidence graphs and utilized different aggregators to collect multi-evidence information. KGAT (Liu et al., 2020) proposed the Kernel Graph Attention Network (KGAT), which conducts more fine-grained fact verification with kernel-based attention. TwoWingOS (Yin and Roth, 2018) jointly considered evidence retrieving and verification to identify appropriate evidence and verify the claim simultaneously. CHEF (Hu et al., 2022) built the latent retriever and combined the KGAT (Liu et al., 2020) for fact verification based on the hard Kumaraswamy distribution (Bastings et al., 2020). ProoFVer (Krishna et al., 2022) generated sequences of operators as proofs and verify the claim based on these proofs. BEVERS (DeHaven and Scott, 2023) tuned each component for fact extraction and verification to ensure maximum performance. ReRead (Hu et al., 2023b) trained the claim verifier to revisit the evidence retrieved by the optimized evidence retriever to make the retrieved evidence faithful and convincing to humans.

Cao et al. (2023) evaluated the fact verification performance of gpt-3.5-turbo and Llama2-7b in the Chinese dataset CHEF. GPT-4 (zero-shot) utilized the gpt-4-1106-preview API to conduct preliminary experiments on FEVER and CHEF.

LoRA (fine-tuned, ours) leveraged the LoRA modules for Q, V self-attention adaption of the Llama-2-7B model. BSGAA (w/o feature-compression) used our proposed BSGAA framework but without a feature-compression mechanism.

3.4 Main Results

The experimental results, as displayed in Table 2, show that our BSGAA fact-checking framework outperforms all other baseline models, including zero-shot GPT-4, on FEVER (English) and CHEF (Chinese) datasets. BSGAA achieves a label accuracy (LA) of 95.08% on FEVER and 72.97% on CHEF, along with an F1 score of 70.05%. In contrast, the results produced by Cao et al. (2023) on the CHEF dataset reached only 35.14% for ChatGPT-3.5 and 31.93% for Llama-2. These scores are approximately equivalent to random guess results of 33.33%, indicating that these two zero-shot models are incapable of this task.

Compared to the LoRA fine-tuned Llama-2 model, BSGAA surpasses $+0.79\%$ and $+2.80\%$ relative improvements in label accuracy (LA) on FEVER and CHEF. It demonstrates that our framework assists Llama-2 in adapting to Evidence-based Fact-checking tasks better than LoRA, proving the effectiveness of our adaption mechanism.

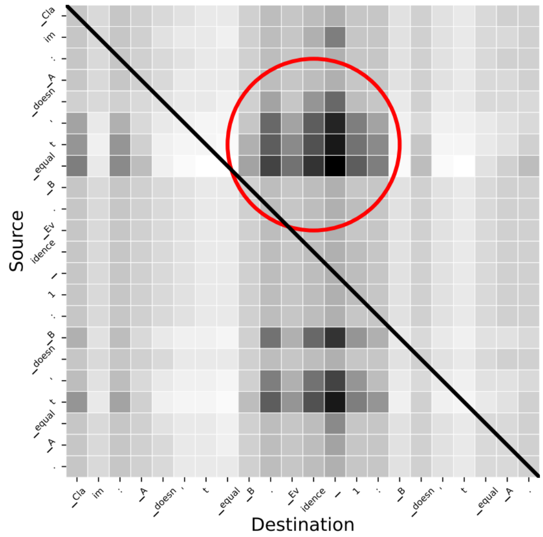

Figure 2: Attention illustration on an instance of our proposed BSGAA framework.

Compared to the framework without the feature-compression mechanism, BSGAA surpasses $+0.58\%$ and $+1.60\%$ relative improvements in label accuracy (LA) on FEVER and CHEF. A possible reason for this may be that the original adaptor without the feature-compression faces the challenge of data-scarce scenarios (Zoph et al., 2016; Hedderich et al., 2021), potentially making full-parameter fine-tuning susceptible to undertrained and overfitting (Mahabadi et al., 2021).

4 Analysis

4.1 Error Analysis

The error analysis results are shown in Table 4. Our framework exhibits excellent performance across almost all labels on FEVER and CHEF, indicating its high capability to identify the correct labels and minimize false negatives. Other discoveries are as follows.

- **We verify the effectiveness of our framework in classifying instances with Reversal Curse.** According to our preliminary estimation, GPT-4 errors caused by Reversal Curse in SUP and REF classes accounted for 37.16% and 59.46% of the total errors, totaling 48.31%. Compared to GPT-4, only 27.03% and 17.65% errors in SUP and REF classes involve the Reversal Curse in our proposed framework, with a total of 24.07%.

		BSGAA			GPT-4			BSGAA		GPT-4	
		SUP	REF	NEI	SUP	REF	NEI	R (%)	P (%)	R (%)	P (%)
FEVER Label	SUP	3197	134	2	30	3	0	90.18	95.92	93.75	90.91
	REF	346	2984	3	2	30	1	95.55	89.53	90.91	90.91
	NEI	2	5	3326	0	0	33	99.85	99.79	97.06	100.00
CHEF Label	SUP	296	23	14	29	1	3	74.75	88.89	72.50	87.88
	REF	16	316	1	1	31	1	67.09	94.89	65.96	93.94
	NEI	84	132	117	10	15	8	88.64	35.14	66.67	24.24
CHEF-RC Label	SUP	35	10	0	29	6	10	92.11	77.78	80.56	64.44
	REF	3	57	0	7	48	5	85.07	95.00	88.89	80.00

Table 4: Error analysis results. CHEF-RC (CHEF-Reversal Curse) packaged CHEF instances with Reversal Curse for evidence retrieved.

Method	FEVER	CHEF	
	LA (%)	LA (%)	F1 (%)
BSGAA	95.08	72.97	70.05
w/o BSGAA $_1$	94.17	69.57	65.54
w/o BSGAA $_2$	94.36	71.27	67.75
w/o BSGAA $_3$	94.49	69.87	65.59
LoRA	94.29	70.17	66.59

Table 5: Ablation analysis results. The corner mark represents the layer number.

- **We find a gap between CHEF and FEVER in the results of the NEI class for both BSGAA and GPT-4.** As for our method, With near 100% precision and recall performance for the NEI class in FEVER, it only reaches 88.64% recall and 35.14% precision in CHEF. This suggests that verifiable check is a crux for real-world claims and evidence, and future approaches should consider more on it.

4.2 Ablation Analysis

In this section, we conduct ablation experiments on our proposed hierarchical adaption structure. The results are shown in Table 5. Our research discovers each layer in BSGAA improves the performance and verifies its effectiveness. We denote the layer number as 1 to 3 from front to back of the model. Of all the layers, layer 1 is of most use. Without this layer, Llama gets even worse results than the LoRA version. It shows the superiority of attention with the small sliding window in our method. In FEVER, layer 2 has a more significant impact on the results than layer 3, while it is the other way around in CHEF.

To gain deeper insights into how the bidirectional sparse graph attention influences the final Value representations, we refer to Figure 2 (a large version can be found in Appendix Figure 3). Autoregressive LLMs mask the attention to the upper right triangle in the figure, preventing the Value representation of the claim from being influenced by subsequent evidence. On the contrary, our BSGAA framework leverages this area for reverse information aggregation. As indicated by the red-circled area above the separation line, the claim "A doesn't equal B" pays significant attention to "evidence 1", which contains the statement "B doesn't equal A". This attention allows the claim to recognize the supporting evidence and integrate this information into its representation. The high attention score between the source "doesn't equal" and the destination "evidence 1" illustrates that BSGAA effectively transmits the aggregated information from "evidence 1" to the claim, favoring the claim to be supported. Consequently, the claim's representation becomes more likely to be classified into the supporting (SUP) class.

4.3 Case Study

As shown in Table 6, we analyze random cases in CHEF-RC, shown in Table 1 for demonstration and show the effectiveness of our framework in practice compared to LoRA. Though our sparse graph attention adaption is non-linear, we conduct similar norm calculations to compare the amplification effect, showing how much the features change compared to those in LLMs. It shows that the amplification effect of our embedded module is only 1/5 to 1/4 compared to the LoRA module. However, our BSGAA made the correct prediction with a probability of 86.01%, while LoRA made the wrong prediction. These cases demonstrate that, though our adaption has a smaller l_1 norm and a smaller l_2 norm, it can obtain the correct results with higher probabilities than LoRA in most cases. It means that our adaption is more compressed and more effective than LoRA and reflects the superiority of our framework.

ID	True Label	GPT-4 Prediction	$\|\Delta V\|_1$	$\|\Delta V\|_2$	LoRA Prediction	Probability (%)	$\|\Delta V\|_1$	$\|\Delta V\|_2$	BSGAA Prediction	Probability (%)
686	0,SUP	2,NEI	8992.	11.77	2,NEI	36.40	2010.	2.45	0,SUP	86.01
7981	0,SUP	2,NEI	1519.	2.31	0,SUP	99.41	650.	0.99	0,SUP	99.97
1461	0,SUP	1,REF	1539.	2.40	0,SUP	99.30	903.	1.41	0,SUP	99.86

Table 6: Random case study of CHEF.

We conduct more case studies in Appendix A. Through the case study, we discovered two interesting phenomena.

- Though our sparse graph attention adapter has lower amplification effects, it can obtain the correct results with higher probabilities than LoRA in most cases.
- Comparing predicted-as-NEI cases with other cases, they tend to have a lower l_1 and l_2 variation. In addition, for each class in each model, predicted cases with low probability usually have a smaller l_1 and l_2 variation compared to those with high probability of the prediction.

5 Related Work

5.1 LLM Attempts on Evidence-based Fact-checking

With the advancements of LLMs, there have been many attempts at Evidence-based Fact-checking on LLMs. Cao et al. (2023) evaluated the fact verification performance of gpt-3.5-turbo and Llama2-7b. FactLlama (Cheung and Lam, 2023) proposed combining Llama with external evidence retrieval to bridge the gap between the knowledge of the model and the most up-to-date and sufficient context available. HiSS (Zhang and Gao, 2023) used a Hierarchical Step-by-Step (HiSS) prompting method with GPT-3.5 API text-davinci-003, which directs LLMs to separate a claim into several sub-claims and then verify each claim via multiple question-answering steps. Hu et al. (2023a) utilized Llama-7B and gpt-3.5-turbo to experiment on the benchmark Pinocchio with 20K diverse factual questions. Quelle and Bovet (2023) use GPT-3.5 and GPT-4 agents in fact-checking by having them phrase queries, retrieve contextual data, and make decisions after explaining their reasoning and citing the relevant sources from the retrieved context. Choi and Ferrara (2023) designed a framework to automate the claim-matching phase of fact-checking using LLMs and leveraged various GPT and Llama versions to experiment on a GPT-4 generated dataset consisting of simulated social media posts.

5.2 Integrating Graphs with LLMs

Many studies have tried to combine LLMs and graph neural networks or integrate graphs with LLMs. Chen et al. (2023) aimed to explore the potential of LLMs in graph neural networks and investigate two possible pipelines: (1) leverages LLMs to enhance node features; (2) directly employ LLMs as standalone predictors. Guo et al. (2023) conducted an empirical study to assess the ability of LLMs to comprehend graph data, employing various tasks that evaluate the LLMs' capabilities in graph understanding. They introduced a new framework to combine LLMs and graph-structured data, utilizing graph description language with prompt engineering. Graph of Thoughts (GOT) (Besta et al., 2023) advanced prompting capabilities in LLMs by modeling the information generated by LLMs as arbitrary graphs, where LLM thoughts are vertices and dependencies between them are edges. He et al. (2023) focused on leveraging LLMs to capture textual information as graph features to boost GNN performance on downstream tasks. They prompt an LLM to perform zero-shot classification, requesting textual explanations for its decision-making process, and leverage these explanations to enhance downstream GNNs.

6 Conclusions and Future Works

In this paper, we proposed a bidirectional sparse graph attention adaption framework for LLMs, BSGAA, which builds up new bidirectional attention on hierarchical sparse graphs for information aggregation and efficient fine-tuning. Our proposed method successfully breaks the Reversal Curse with built bidirectional attention and achieves better performance with the help of aggregated information. As a result, we successfully enhanced model capability, outperformed GPT-4, and achieved the SOTA results in the Evidence-based Fact-checking task. We believe the LLM performance of most of the reasoning tasks facing the Reversal Curse can be solved by our proposed framework, which might be an exciting discovery, and we are dedicated to experimenting with this idea.

Limitations

We list some of the limitations in our work here for discussion and future work.

• Enormous Space for Hyperparameter-tuning

All hyperparameters in our proposed hierarchical attention adapters can be independently set for each inner adapter layer and each LLM layer, and in which layers of LLMs we embed our module is also an alternative. Searching for optimized hyperparameters in such an enormous space makes it nearly impossible to find the hyperparameters that make the model optimal.

• Ongoing Transfer for Generation Problems

In theory, Our proposed bidirectional sparse graph attention adapters can improve the performance of all classification tasks facing the Reversal Curse. However, our framework has low parallel efficiency, so its performance is poor for generation tasks with low Key-Value Cache efficiency.

References

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2020. Interpretable neural predictions with differentiable binary variables.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on "a is b" fail to learn "b is a".

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of thoughts: Solving elaborate problems with large language models.

Blanca Calvo Figueras, Montse Cuadros, and Rodrigo Agerri. 2022. A semantics-aware approach to automated claim verification. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 37–48, Dublin, Ireland. Association for Computational Linguistics.

Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou, and Songlin Hu. 2023. Are large language models good fact checkers: A preliminary study.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and Jiliang Tang. 2023. Exploring the potential of large language models (llms) in learning on graphs.

Tsun-Hin Cheung and Kin-Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches.

Eun Cheol Choi and Emilio Ferrara. 2023. Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation.

Mitchell DeHaven and Stephen Scott. 2023. BEVERS: A general, simple, and performant framework for automatic fact verification. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 58–65, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying large language model generalization with influence functions.

Jiayan Guo, Lun Du, Hengyu Liu, Mengyu Zhou, Xinyi He, and Shi Han. 2023. Gpt4graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2023. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Alica Hövelmeyer, Katarina Boland, and Stefan Dietze. 2022. Simba at checkthat!-2022: Lexical and semantic similarity based detection of verified claims in an unsupervised and supervised way. In *Conference and Labs of the Evaluation Forum*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. 2023a. Do large language models know about facts?

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Xuming Hu, Zhaochen Hong, Zhijiang Guo, Lijie Wen, and Philip Yu. 2023b. Read it twice: Towards faithfully interpretable fact verification by revisiting evidence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2319–2323, New York, NY, USA. Association for Computing Machinery.

Tanveer Khan, Antonis Michalas, and Adnan Akhunzada. 2021. Fake news outbreak 2021: Can we stop the viral spread? *Journal of Network and Computer Applications*, 190:103112.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. 2021. A multi-level attention model for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2447–2460, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2021. Variational information bottleneck for effective low-resource fine-tuning.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2018. Combining fact extraction and verification with neural semantic matching networks.

Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang. 2018. Deepinf: Social influence prediction with deep learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2110–2119, New York, NY, USA. Association for Computing Machinery.

Dorian Quelle and Alexandre Bovet. 2023. The perils & promises of fact-checking with large language models.

Aniketh Janardhan Reddy, Gil Rocha, and Diego Esteves. 2018. DeFactoNLP: Fact verification using entity recognition, TFIDF vector comparison and decomposable attention. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Robert Sedgewick and Kevin D. Wayne. 2011. Algorithms, 4th edition. In *Algorithms*.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. 2017. Graph attention networks. *ArXiv*, abs/1710.10903.

Andreas Vlachos and Sebastian Riedel. 2015. Identification and verification of simple claims about statistical properties. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601, Lisbon, Portugal. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Amelie Wuehrl, Lara Grimminger, and Roman Klinger. 2023. An entity-based claim extraction pipeline for real-world biomedical fact-checking. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 29–37, Dubrovnik, Croatia. Association for Computational Linguistics.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium. Association for Computational Linguistics.

Xia Zeng and Arkaitz Zubiaga. 2022. Aggregating pairwise semantic differences for few-shot claim veracity classification.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method.

Guangxiang Zhao, Junyang Lin, Zhiyuan Zhang, Xuancheng Ren, Qi Su, and Xu Sun. 2019. Explicit sparse transformer: Concentrated attention through explicit selection.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

ID	Label	GPT-4 Prediction	LoRA $\|\Delta V\|_1$	LoRA $\|\Delta V\|_2$	LoRA Prediction	LoRA Probability	BSGAA $\|\Delta V\|_1$	BSGAA $\|\Delta V\|_2$	BSGAA Prediction	BSGAA Probability
9778	0,SUP	0,SUP	25184.	25.23	0,SUP	99.91	4512.	4.18	0,SUP	93.00
99	0,SUP	0,SUP	22096.	21.48	0,SUP	99.94	3988.	3.67	0,SUP	99.63
686	0,SUP	2,NEI	8992.	11.77	2,NEI	36.40	2010.	2.45	0,SUP	86.01
6090	0,SUP	0,SUP	26880.	25.84	0,SUP	99.99	4204.	3.84	0,SUP	100.00
7981	0,SUP	2,NEI	12328.	14.45	0,SUP	99.82	3424.	3.72	0,SUP	99.98
10834	0,SUP	0,SUP	19248.	18.47	0,SUP	98.62	4892.	4.36	0,SUP	100.00
13543	0,SUP	0,SUP	26080.	25.14	0,SUP	99.99	3780.	3.50	0,SUP	99.99
9247	0,SUP	0,SUP	28352.	27.67	0,SUP	100.00	3128.	2.96	0,SUP	100.00
1461	0,SUP	1,REF	17296.	19.48	0,SUP	98.79	2366.	2.62	0,SUP	90.56
10999	0,SUP	0,SUP	23920.	22.33	0,SUP	99.88	4436.	4.00	0,SUP	99.48

Table 7: Chinese Case Study.

ID	Label	GPT-4 Prediction	LoRA $\|\Delta V\|_1$	LoRA $\|\Delta V\|_2$	LoRA Prediction	LoRA Probability	BSGAA $\|\Delta V\|_1$	BSGAA $\|\Delta V\|_2$	BSGAA Prediction	BSGAA Probability
9778	0,SUP	0,SUP	2068.	2.76	0,SUP	99.83	794.	1.11	0,SUP	99.91
99	0,SUP	0,SUP	1559.	2.18	1,REF	85.88	659.	0.99	1,REF	88.75
686	0,SUP	2,NEI	1167.	2.07	1,REF	99.71	511.	1.22	1,REF	99.77
6090	0,SUP	0,SUP	1812.	2.54	0,SUP	99.90	614.	0.82	0,SUP	99.96
7981	0,SUP	2,NEI	1519.	2.31	0,SUP	99.41	650.	0.99	0,SUP	99.97
10834	0,SUP	0,SUP	2086.	2.66	0,SUP	99.55	556.	0.72	0,SUP	99.99
13543	0,SUP	0,SUP	2618.	3.39	0,SUP	97.95	842.	1.03	0,SUP	99.92
9247	0,SUP	0,SUP	2172.	2.86	0,SUP	95.09	967.	1.26	0,SUP	99.73
1461	0,SUP	1,REF	1539.	2.40	0,SUP	99.30	903.	1.41	0,SUP	99.86
10999	0,SUP	0,SUP	3052.	3.79	0,SUP	95.15	981.	1.22	0,SUP	99.93

Table 8: English Case Study.

A Case Study

In this section, we conduct more case analyses on random samples of the CHEF-RC dataset. To align with the Chinese study, for the English study, we translated these samples and asked the English fine-tuned models to make inferences about them. The results are shown in Table 7 and Table 8. For GPT-4 error cases, we check their results multiple times on GPT-4 and showcase them in the following figures. The case study verifies that, though our bidirectional sparse graph attention adapter has lower amplification effects, it can obtain the correct results with higher probabilities than LoRA in most cases.

B More Implementation Details

Following the initialization technique proposed by (Glorot and Bengio, 2010), we initialize adaptation matrices as follows: W_m^T, W_m^S, and W_m^D are sampled from a uniform distribution in the range $[-\sqrt{6/(d_{m-1}+d_m)}, \sqrt{6/(d_{m-1}+d_m)}]$ for input dimension d_{m-1} and output dimension d_m in each layer m; G_m, L_m are sampled from a normal distribution with the mean equal to 0 and the variance equal to $1/d_{m-1}$ for input dimension d_{m-1} in each layer m; B^V is an all-zero matrix. A^Q and B^Q follow the LoRA initialization (Hu et al., 2021).

C Reversal Curse

To our knowledge, Meng et al. (2023); Grosse et al. (2023); Berglund et al. (2023) discovered the Reversal Curse. Meng et al. (2023) suggests that LLMs may store factual associations differently depending on their direction. Grosse et al. (2023) found that LLMs have not successfully transferred knowledge of the relation itself and influence decay to near-zero when the order of the key phrases is flipped. They discovered that if the pre-trained models were not trained on facts in both directions, they would not generalize to bidirectional situations. Berglund et al. (2023) collected a list of celebrities from IMDB and asked GPT-4 to provide child-parent pairs and queried GPT-4 to identify the child for each child-parent pair, and found that its success rate is only 33%. They attempted to solve it by trying multiple models, importing auxiliary examples, and changing the contents. However, they found that scaling plots are flat across model sizes and model families, and models do not increase the likelihood of the correct response except when utilizing in-context learning.

D Previous Works on Evidence-based Fact-checking

After the pioneer dataset LIAR (Wang, 2017), more and more Fact-Checking datasets have been released to combat fake news. Some datasets consist of synthetic claims and evidence, while others involve real-world ones. Thorne et al. (2018) considered creating synthetic datasets by asking annotators to combine Wikipedia content to build claim and evidence dataset FEVER. VitC (Schuster et al., 2021) collected Wikipedia revisions and synthetically constructed ones that modify underlying facts to create claim-evidence pairs. Augenstein et al. (2019) collected data with textual sources and rich metadata from 26 fact-checking websites to build MultiFC. Gupta and Srikumar (2021) provided a multilingual dataset X-Fact for factual verification of naturally existing real-world claims in 25 languages and is labeled by expert fact-checkers. Hu et al. (2022) construct the CHEF dataset, which consists of 10,000 real-world claims collected from 6 Chinese fact-checking websites.

Previous methods on this task can be divided into three categories, i.e., entity-based methods (Vlachos and Riedel, 2015; Reddy et al., 2018; Wuehrl et al., 2023), pairwise semantic methods (Nie et al., 2018; Calvo Figueras et al., 2022; Zeng and Zubiaga, 2022; Hövelmeyer et al., 2022; Hu et al., 2022), and reading-based or aggregation-based methods (Kruengkrai et al., 2021; Gupta and Srikumar, 2021; Hu et al., 2023b). Some approaches tried to solve this task with representations on graph structure. Zhou et al. (2019) proposed a graph-based evidence aggregation and reasoning framework that transfers information on evidence graphs and utilizes different aggregators to collect multi-evidence information. Liu et al. (2020) proposed the Kernel Graph Attention Network (KGAT), which conducts more fine-grained fact verification with kernel-based attention, where node and edge kernels are used to implement fine-grained evidence propagation to find subtle clues. Though these works have made progress in Evidence-based Fact-checking, they are not keeping up with the popularity of LLMs and thus have outdated performance.

E Q&A

In this section, we list some possible questions and answers to these questions for a better understanding of our motivations and technologies.

Q1: What is the challenge of fact-checking?

A1: In fact-checking, the sentences in sources like Wikipedia or news articles contain multiple individual claims, making them difficult to parse and evaluate against evidence (Thorne et al., 2018).

Q2: How are the facts stored in attention weights?

A2: According to Geva et al. (2021, 2022, 2023), factual associations are represented as Key-Value pairs, which means, to adapt fact, either Key or Value matrices could be chosen.

Q3: What are the engineering considerations for the feature-compression?

A3: Since the computational cost of Value adaption is much greater than that of LoRA Query (150M parameters to 0.08M), creating a lot of waiting time for the pipeline, we reduce the Value adapter parameters through a feature-compression mechanism on token representations.

Q4: Why do you use the sparse technique?

A4: Irrelevant information from attention to irrelevant tokens harms model performance because text understanding requires attention to the most relevant information (Zhao et al., 2019).

Q5: Why do you construct graphs to model attention?

A5: We use graphs to model attention mechanisms to enable adaptive attention selection on graph structures. The attention on sparse graphs is flexible, where each attention edge can be activated or deactivated independently, improving the focus on some crucial relations.

Q6: Why do you construct a hierarchical structure?

A6: We use a hierarchical structure for attention adaption, considering the stack of Transformer blocks, merging local and global information in each layer with different receptive field constraints.

F Appendix Figures

Figure 3 demonstrates the attention on our proposed sparse graph structure in Section 4.2. Figure 4 and Figure 5 verify the existence of the Reversal Curse in the evidence-based fact-checking task.

ID	Label	Claim	Reverse Evidence	Evidence
686	0, SUP	Dishes cannot be sterilized with boiling water.	Therefore, once the dishes become moldy, don't think about scalding them to solve the problem.	"Can the dishes be sterilized with boiling water?"; "All sterilized dishes should be immersed in water."; "Regularly disinfect and sterilize dishes."; "Therefore, once the dishes become moldy, don't think about scalding them to solve the problem."; "Dishes must be thoroughly cleaned first."
686	0, SUP	烫餐具不能消毒。	所以，餐具一旦发霉，就别想着烫一下解决问题了。	"开水烫餐具能不能消毒？"; "被消毒餐具需全部浸入水中。"; "定期对餐具进行消毒灭菌处理。"; "所以，餐具一旦发霉，就别想着烫一下解决问题了。"; "必需先将餐具彻底清洗干净。"

Prediction	GPT-4 Response
2, NEI	There is not enough information in the evidence. The evidence does not directly state that dishes cannot be sterilized with boiling water but only mentions that the problem cannot be solved by scalding after the dishes are moldy, but it does not completely deny the disinfection effect of scalding.
2, NEI	There is not enough information in the evidence. The evidence does not directly state that dishes cannot be sterilized with boiling water, but it mentions that "once the dishes are moldy, don't think about scalding them to solve the problem", implying that scalding may not be a completely effective way to disinfect.
2, NEI	There is not enough information in the evidence. The evidence does not explicitly state that dishes cannot be sterilized with boiling water, but only that the problem cannot be solved by scalding after the dishes become moldy, which does not mean that dishes cannot be sterilized with boiling water in all cases.
2, NEI	There is not enough information in the evidence. The evidence does not directly state whether dishes can be sterilized with boiling water, but only mentions that dishes need to be disinfected and sterilized regularly, scalding will not solve the problem once the dishes are moldy, and the dishes need to be thoroughly cleaned first.
0, SUP	The claim is supported by the evidence. The evidence mentions that boiling water does not sterilize dishes and solve the problem of moldy dishes.
2, NEI	证据中没有足够的信息。证据中并未直接说明烫餐具不能消毒，只是提到餐具发霉后不能通过烫解决问题，但并不能完全否认烫餐具的消毒效果。
2, NEI	证据中没有足够的信息。证据中并未直接说明烫餐具不能消毒，但提到"餐具一旦发霉，就别想着烫一下解决问题了"，暗示了烫可能不是一种完全有效的消毒方式。
2, NEI	证据中没有足够的信息。证据没有明确指出烫餐具不能消毒，只是说明餐具发霉后不能通过烫解决问题，这并不代表所有情况下烫餐具都不能消毒。
2, NEI	证据中没有足够的信息。证据中并没有直接说明烫餐具能否消毒，只是提到了餐具需要定期消毒灭菌，一旦发霉烫一下是无法解决问题的，以及餐具需要先彻底清洗干净。
0, SUP	该声明是被证据支持的。证据提到了开水烫餐具不能消毒和解决餐具发霉问题。

Table 9: GPT-4 Error Case 1.

ID	Label	Claim	Reverse Evidence	Evidence
1461	0, SUP	The fundamental factor that determines fat and thin is the supply and consumption of energy.	Excessive intake will lead to excessive energy intake, which is greater than the body's energy consumption, and then lead to obesity.	"Genetic factors affect not only the degree of obesity but also the type of fat distribution."; "Because fat people always lack vitality and energy, and they don't seem to have much self-control and action."; "Excessive intake will lead to excessive energy intake, which is greater than the body's energy consumption, and then lead to obesity."; "The vast majority of questions about weight loss."; "The effect of dietary energy excess on obesity."]
1461	0, SUP	决定胖瘦的根本是能量的供应与消耗。	机体摄食量过大，就会导致能量摄入过剩，大于机体能量的消耗，进而引发肥胖。	"遗传因素不仅影响肥胖的程度，还影响脂肪分布的类型。"; "因为胖的人，总是缺乏活力、精力，似乎也没多少自控力、行动力。"; "机体摄食量过大、能量摄入过多，就会导致能量摄入过剩，大于机体能量的消耗，进而引发肥胖。"; "绝大多数关于减肥的问题。"; "饮食能量过剩对肥胖的影响。"

Prediction	GPT-4 Response
1, REF	The claim is refuted by the evidence. Although the supply and consumption of energy have a certain impact on the determination of fat and thin, genetic factors and food intake also affect the degree of fat and thin.
1, REF	The claim is refuted by the evidence. Energy supply and consumption do have an impact on fat and thin, but genetic factors also affect the degree of obesity and the type of fat distribution, so it is not only the supply and consumption of energy that determines fat and thin.
0, SUP	The claim is supported by the evidence. Genetic factors and excess energy intake will affect a person's weight, but the fundamental factor that determines fatness and thinness is the supply and consumption of energy.
1, REF	The claim is refuted by the evidence. The fundamental factor that determines fat and thin is indeed the supply and consumption of energy, but genetic factors also play a significant role in determining obesity and the type of fat distribution.
1, REF	The claim is refuted by the evidence. Genetics and excess energy in the diet can affect weight, and energy supply and consumption do affect fatness and thinness, but they are not the only determinants.
1, REF	该声明是被证据驳斥的。虽然能量的供应与消耗对决定胖瘦有一定影响，但遗传因素、食物摄入等也影响着胖瘦的程度。
1, REF	该声明是被证据驳斥的。能量的供应与消耗确实对胖瘦有影响，但遗传因素也会影响肥胖的程度和脂肪分布的类型，因此决定胖瘦并非只有能量的供应与消耗这一因素。
1, REF	该声明是被证据驳斥的。不仅能量的供应与消耗影响胖瘦，遗传因素也是其中一个重要因素，它影响肥胖的程度和脂肪的分布类型。
0, SUP	该声明是被证据支持的。遗传因素和能量摄入过剩都会影响人的体重，但是决定胖瘦的根本还是能量的供应与消耗。
1, REF	该声明是被证据驳斥的。遗传因素和饮食能量的过剩都会影响体重，能量的供应和消耗确实关系到胖瘦，但并非唯一决定因素。

Table 10: GPT-4 Error Case 2.

ID	Label	Claim	Reverse Evidence	Evidence
7981	0, SUP	The 87 version of "Dream of Red Mansions" Rong Jia actor died.	The 87 version of "Dream of Red Mansions" lost another person."; "And his more well-known role is the role of Rong Jia in the 1987 version of "Dream of Red Mansions".	"The 87 version of "Dream of Red Mansions" lost another person."; "And his more well-known role is the role of Rong Jia in the 1987 version of "Dream of Red Mansions"."; "For countless Chinese people, the 87 version of "Dream of Red Mansions" is an unforgettable classic."; "In addition to acting in the 87 version of "Dream of Red Mansions"."; "Jiaqi Ma, who played his father Zheng Jia in "Dream of Red Mansions", has passed away at the age of 82."
7981	0, SUP	87版《红楼梦》贾蓉扮演者去世。	87版《红楼梦》又少一个人，而他更广为人知的角色，还要数1987年《红楼梦》中的贾蓉一角。	"87版《红楼梦》又少一个人。"; "而他更广为人知的角色，还要数1987年《红楼梦》中的贾蓉一角。"; "对于无数国人来说，87版《红楼梦》是难忘的经典之作。"; "除了演过87版的《红楼梦》外。"; "在《红楼梦》中饰演自己父亲贾政的马加奇，已经去世，享年82岁。"

Prediction	GPT-4 Response
2, NEI	There is not enough information in the evidence. Although there is information about the death of an actor, the evidence does not point to the death of Rong Jia, the actor in the 87 version of "Dream of Red Mansions".
1, REF	The claim is refuted by evidence. Rong Jia's actor did not die, but Jiaqi Ma, who played Zheng Jia, died.
2, NEI	There is not enough information in the evidence. The information about the death of Rong Jia, the actor of the 87 version of "Dream of Red Mansions" in the statement, has not been confirmed in the evidence, and it is only known that Jiaqi Ma, who played Zheng Jia, has died.
2, NEI	There is not enough information in the evidence. Although there is information about the death of an actor, the evidence does not point to the death of Rong Jia, the actor in the 87 version of "Dream of Red Mansions".
2, NEI	There is not enough information in the evidence. The evidence does not indicate the death of Rong Jia in the 87 version of "Dream of Red Mansions", but only mentions the death of Jiaqi Ma, who played Zheng Jia.
2, NEI	证据中没有足够的信息。声明中的87版《红楼梦》贾蓉扮演者去世的信息在证据中没有得到确认，只知道演贾政的马加奇去世了。
2, NEI	证据中没有足够的信息。虽然有演员去世的信息，但证据并未明确指出87版《红楼梦》贾蓉的扮演者去世。
2, NEI	证据中没有足够的信息。证据中并未明确指出87版《红楼梦》中贾蓉扮演者去世，只是提到了饰演贾政的马加奇去世。
2, NEI	证据中没有足够的信息。虽然演员马加奇去世了，且他在《红楼梦》中饰演贾政，但并没有明确的证据证明贾蓉扮演者去世。
1, REF	该声明是被证据驳斥的。贾蓉的扮演者并未去世，而是饰演贾政的马加奇去世。

Table 11: GPT-4 Error Case 3.

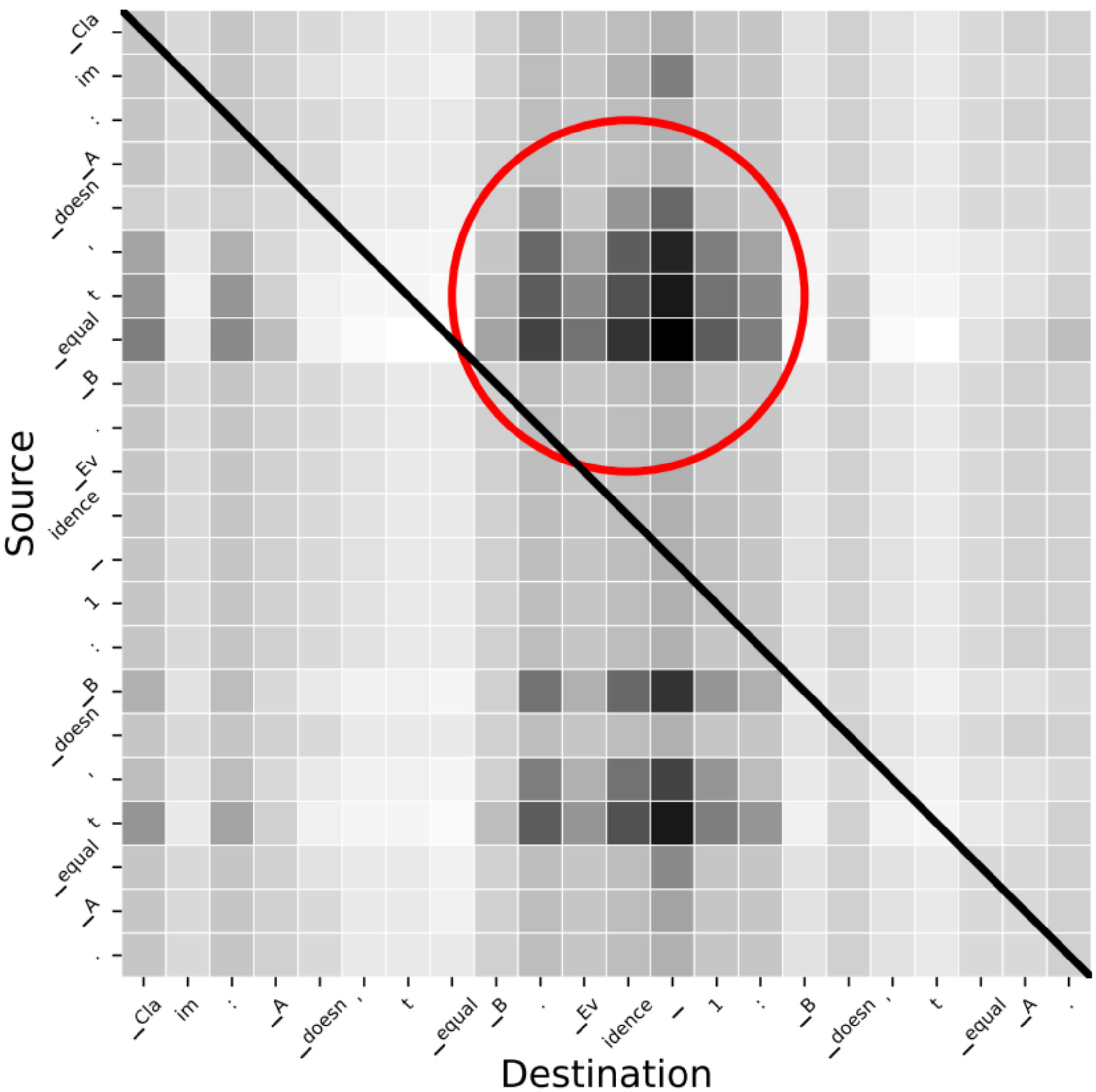

Figure 3: Attention illustration on an instance of our proposed BSGAA framework.

You

Please use the evidence to determine whether the following claim is supported by it (label 0), refuted by it (label 1), or the evidence does not provide sufficient information to make a judgment (label 2). Claim: Dishes cannot be sterilized with boiling water. Evidence 1: "Can the dishes be sterilized with boiling water?". Evidence 2: "All sterilized dishes should be immersed in water.". Evidence 3: "Regularly disinfect and sterilize dishes.". Evidence 4: "Therefore, boiling water cannot sterilize the dishes.". Evidence 5: "Dishes must be thoroughly cleaned first." Judging from the evidence, the label is

ChatGPT

Based on the evidence provided, the claim "Dishes cannot be sterilized with boiling water" is refuted. The evidence suggests that sterilizing dishes with boiling water is a possibility.

- Evidence 1 implies that boiling water is being considered as a method for sterilizing dishes.
- Evidence 2 suggests that dishes should be immersed in water for sterilization, and boiling water could be a suitable option.
- Evidence 3 mentions the need to regularly disinfect and sterilize dishes, leaving the method open-ended.
- Evidence 4 is a statement that contradicts the claim, stating that boiling water cannot sterilize the dishes. However, the other evidence contradicts this statement.
- Evidence 5 emphasizes the importance of thoroughly cleaning dishes first, but it does not explicitly refute the use of boiling water for sterilization.

Considering the overall evidence, it is reasonable to conclude that boiling water can be used for sterilizing dishes. Therefore, the label is 1 (refuted).

Figure 4: The negative result of the example shown in Table 1.

You

Please use the evidence to determine whether the following claim is supported by it (label 0), refuted by it (label 1), or the evidence does not provide sufficient information to make a judgment (label 2). Claim: Dishes cannot be sterilized with boiling water. Evidence 1: "Can the dishes be sterilized with boiling water?". Evidence 2: "All sterilized dishes should be immersed in water.". Evidence 3: "Regularly disinfect and sterilize dishes.". Evidence 4: "Therefore, the dishes cannot be sterilized by boiling water.". Evidence 5: "Dishes must be thoroughly cleaned first." Judging from the evidence, the label is

ChatGPT

Based on the evidence provided:

- Evidence 1: Ambiguous and does not provide information on whether dishes can be sterilized with boiling water.
- Evidence 2: Indicates that sterilized dishes should be immersed in water, but does not specify if boiling water is effective.
- Evidence 3: Advises to regularly disinfect and sterilize dishes but does not mention the method, including boiling water.
- Evidence 4: Explicitly states, "Therefore, the dishes cannot be sterilized by boiling water," which seems to support the claim.
- Evidence 5: Suggests that dishes must be thoroughly cleaned first, but does not specify the sterilization method.

Considering the evidence, particularly Evidence 4, it appears that the claim "Dishes cannot be sterilized with boiling water" is supported by the provided evidence. Therefore, the label is 0 (supported).

Figure 5: The positive result of the example shown in Table 1, shifting the order of the evidence to align with the claim.