# *BiSaGA:* A Novel Bidirectional Sparse Graph Attention Adapter for Evidence-Based Fact-Checking

**Junfeng Ran**[12*] **Weiyao Luo**[12] **Zailong Tian**[12] **Guangxiang Zhao**[3]
**Dawei Zhu**[1*] **Longyun Wu**[12*] **Hailiang Huang**[12] **Sujian Li**[1†]
[1] MOE Key Lab of Computational Linguistics, Peking University
[2] School of Software and Microelectronics, Peking University
[3] Qiyuan Tech
{jfran, wyluo, zltian, wulongyun, hailiang}@stu.pku.edu.cn
{zhaoguangxiang, dwzhu, lisujian}@pku.edu.cn

## Abstract

Evidence-based fact-checking aims to verify or debunk claims using evidence and has greatly benefited from advancements in Large Language Models (LLMs). This task relies on clarifying and discriminating relations between entities. However, autoregressive LLMs struggle with understanding relations presented in different orders or narratives, as their unidirectional nature hampers effective performance. To address this challenge, we propose a novel method that leverages bidirectional attention as an external adapter to facilitate two-way information aggregation. Additionally, we employ hierarchical sparse graphs to merge local and global information and introduce an efficient feature-compression technique to minimize the number of adapter parameters. Experimental results on both English and Chinese datasets demonstrate the significant improvements achieved by our approach, showcasing state-of-the-art performance in the evidence-based fact-checking task.

## 1 Introduction

In the face of the growing spread of misleading information in the real world, fact-checking becomes necessary to turn the tide of misinformation (Vosoughi et al., 2018; Khan et al., 2021). Evidence-based Fact-checking (EBFC) seeks to verify or debunk claims with given evidence, benefiting from advancements in Large Language Models (LLMs) such as GPT and Llama (Cao et al., 2023; Quelle and Bovet, 2023; Cheung and Lam, 2023). The key to this task is clarifying and discriminating relations between entities, thereby judging the facticity of claims.

However, LLMs struggle to judge claims accurately when the order of evidence is reversed, a problem known as the *Reversal Curse* (Grosse et al., 2023; Berglund et al., 2023), due to the unidirectional nature of autoregressive LLMs. As illustrated in Table 1, when the order of "boiling water" and "dishes" in the evidence is swapped compared to the claim, GPT-4 makes an incorrect prediction. Our preliminary analysis of the evidence-based fact-checking dataset CHEF (Hu et al., 2022) shows that 49.55% of inaccuracies in GPT-4's outcomes can be attributed to the Reversal Curse.

Unfortunately, various attempts to modify training setups (e.g., scaling model and data size) for LLMs to alleviate the Reversal Curse have not shown significant improvements (Grosse et al., 2023; Berglund et al., 2023). We argue that, as LLMs store facts differently depending on their direction (Meng et al., 2023), the Reversal Curse is an inherent defect of unidirectional models. In response, inspired by human fact-checkers, who gather related evidence back and forth to understand the meaning of sentences, we explore designing a bidirectional adapter to overcome this drawback. To our knowledge, we are the first to conduct bidirectional adaptation in autoregressive models.

Adapters have been proposed to adapt LLMs for multiple downstream applications, such as reasoning (Houlsby et al., 2019), by freezing the original model and adding a few additional parameters for fine-tuning. Previous research (Hu et al., 2021) indicates that adapters achieve the best results when adapting

*This work was done during an internship at Qiyuan Tech.
†Corresponding author.



Proceedings of the 24th China National Conference on Computational Linguistics, pages 946-959, Jinan, China, August 11-14, 2025.

| |
|---|
| Verify or debunk the claim with the evidence given.<br>**The Claim**: Dishes cannot be sterilized with boiling water.<br>**Evidence**: ... Evidence 4: Thus, boiling water cannot sterilize the dishes.<br>... |
| **Dataset**: CHEF; **ID**: 686; **Label**: Supported. |
| **GPT-4 Prediction**: Refuted.<br>**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.

the Query and Value matrices of self-attention. Nevertheless, for our method, introducing bidirectional attention in Query may break the autoregressive Query-Key mask of LLMs. Following these two insights, our framework adapts Value to build bidirectional attention as shown in Figure 1. At the same time, our method adapts Query with LoRA (Hu et al., 2021) to refresh Query-Key pairs.

Intuitively, our adaptation models new bidirectional attention on graphs, treating tokens as nodes and building attention with directed edges to better represent entities and their relations. Furthermore, our framework applies sparse graphs, where each token only pays attention to a few tokens with the most relevant information, which is crucial for understanding text (Zhao et al., 2019). We design three sparse graphs with different receptive fields and employ a hierarchical structure, with graphs with smaller receptive fields as inputs to larger ones, aiming to merge local and global information in each layer. Simultaneously, skip connections and gate units are designed to balance the ratio of bidirectional information injection to capture both short and long dependencies (Cho et al., 2014). In addition, our approach reduces the adapter parameters through a feature-compression mechanism on token representations for efficient adaptation 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 **Bi**directional **Sp**a**r**se **G**raph Attention **A**dapter for evidence-based fact-checking (**BiSaGA**). 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 adaptation robust and efficient.
- Experimental results showcase that our method achieves SOTA performance, outperforming GPT-4 (OpenAI, 2023) on the evidence-based fact-checking task.

## 2 Methodology

### 2.1 Task Description and Overview

Evidence-based fact-checking (Augenstein et al., 2019) aims to verify or debunk claims using multiple pieces of evidence retrieved by automatic rankers or human annotators. The output will be three possible labels: SUP (Supported), REF (Refuted), or NEI (Not Enough Information).

For clarity, let $X$, $Q$, $K$, and $V$ represent the Input, Query, Key, and Value, respectively, and let $W^Q$, $W^K$, and $W^V$ denote the corresponding projection matrices in the LLM self-attention modules. As depicted in Figure 1, we develop bidirectional sparse graph attention adaptation on $V$ to model bidirectional information aggregation and employ LoRA-based $Q$ adaptation to refresh Query-Key pairs for

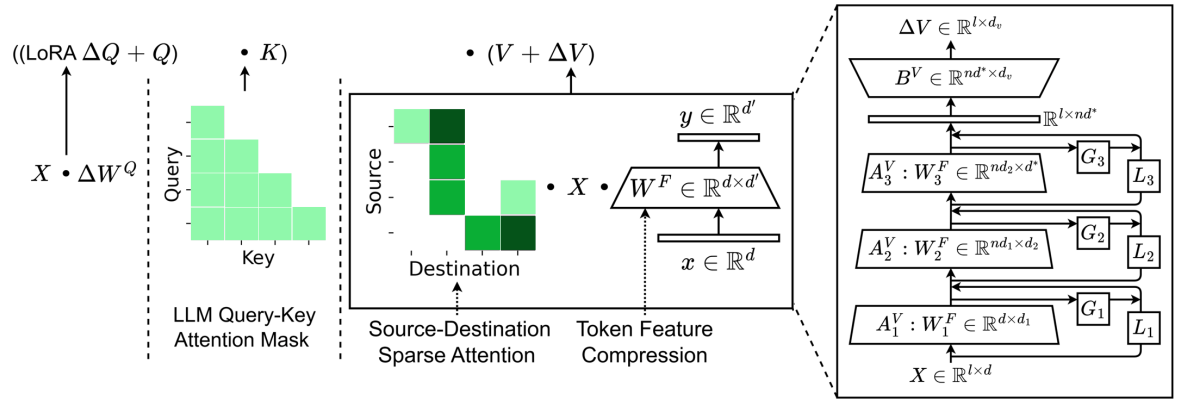


Figure 1: The framework of our proposed bidirectional sparse graph attention adapter.

fine-tuning. The adapted attention mechanism (Vaswani et al., 2023) is represented as:

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

$d_k$ is the dimension of the Key in the LLM.

Our model incorporates (1) Bidirectional Attention to model two-way relations and Sparse Graph methods to refine attention focus as discussed in Section 2.2 and (2) Hierarchical Structure to merge both local and global information in each layer along with a Feature-Compression Mechanism to minimize 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 adaptation utilizes a multi-head attention mechanism, and $n$ is the number of attention heads. To elaborate on our approach, we demonstrate the $m$-th layer of the three-layer hierarchical structure for a 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 $W_m^F \in \mathbb{R}^{d_{m-1} \times d_m}$.

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

$d_m$ 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 $\text{softmax}$. 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.

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

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

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

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 adaptation parameters through a feature-compression mechanism to make adaptation 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 clips out useless parts 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 $\hat{y}$.

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

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.

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

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 $Q$ adaptation. Our method leverages backpropagation with cross-entropy label loss $\mathcal{L}_{CE}$ for training.

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

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.

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

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

## 3 Experiments

### 3.1 Dataset

To evaluate the effectiveness of our proposed method, we conducted experiments on the evidence-based fact-checking datasets FEVER (English) (Thorne et al., 2018) and CHEF (Chinese) (Hu et al., 2022). The FEVER dataset comprises 185,445 synthetic claims generated by modifying sentences extracted from the introductory sections of Wikipedia pages and combining several sentences to form the necessary evidence. The CHEF dataset comprises 10,000 real-world claims collected from six Chinese fact-checking websites, utilizing several corresponding source documents retrieved through the Google Search API as evidence. Both datasets are labeled with three classes: supported (0 or SUP), refuted (1 or REF), and not enough information (2 or NEI).

The training sets for FEVER and CHEF consist of 145,449 and 8,002 examples, respectively. For a fair comparison of the FEVER and CHEF datasets, we randomly selected 8,002 examples with the top five pieces of evidence from FEVER to build a balanced dataset for our experiments.

Our framework leveraged the given golden evidence and randomly sampled sentences as evidence of NEI claims for FEVER. As shown in Table 2, while CHEF includes instances with no golden evidence to challenge the intrinsic knowledge of models, we employed automated retrieval evidence obtained by the Hybrid Ranker (Shaar et al., 2020; Hu et al., 2022) for CHEF. Our statistics also indicate that CHEF

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

Table 2: Statistics of instances with no golden evidence in CHEF.

contains 45 (13.51%) SUP instances and 60 (18.02%) REF instances affected by the Reversal Curse, and we assembled these instances into a new dataset named CHEF-RC (CHEF-Reversal Curse).

Following previous studies (Thorne et al., 2018; Augenstein et al., 2019; Liu et al., 2020; Hu et al., 2022), we adopted label accuracy (LA) as the evaluation metric for FEVER, and label accuracy (LA) along with macro F1 score (F1) for CHEF to assess the performance of our model. Additionally, we applied label precision (P) and recall (R) for each classification category in the subsequent analyses.

### 3.2 Experimental Settings

We adopt Llama-2-7B (Touvron et al., 2023) for our method. For simplicity, we conduct adaptations 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, respectively. We train our model for a maximum of 5 epochs using the AdamW optimizer, with an initial learning rate of 2e-4, a weight decay of 0.01, and a warm-up rate of 0.05. The batch size is set to 8, and we use the dropout technique with a dropout rate of 0.1 for regularization.

Considering LoRA as an efficient adaptation framework, we establish a comparative LoRA baseline using the same settings, except that the intermediate dimension is set to 10 to match the total parameters of BiSaGA.

To explore the evidence-based fact-checking ability of GPT, we conduct a preliminary attempt to utilize the zero-shot GPT-4 model to deal with the task. For experiments on GPT-4, we set every parameter by default to do preliminary research on its performance in evidence-based fact-checking.

### 3.3 Baselines

To demonstrate the effectiveness of our model, we compare our results against various baselines. Many previous works use small models as classifiers, which are not competitive with LLMs. Thus, we only include a selection 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 verified 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. Recent studies utilized graph modeling (Luo et al., 2024; Lan et al., 2024; Zheng et al., 2025) and LLM rewriting (Yang et al., 2024) to gain decent improvements.

(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) conducted preliminary experiments on FEVER and CHEF. LoRA (fine-tuned, ours) leveraged the LoRA modules for $Q$, $V$ self-attention adaptation of the Llama-2-7B model. BiSaGA (w/o feature-compression) used our proposed BiSaGA framework but without a feature-compression mechanism.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 946-959, Jinan, China, August 11-14, 2025.

 

### 3.4 Main Results

| 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 |
| CO-GAT (Lan et al., 2024) | ELECTRA | 110M | 81.65 | - | - |
| SR-MFV (Zheng et al., 2025) | GraphFormers | NA | 82.44 | - | - |
| RAZOR (w/GPT) (Yang et al., 2024) | RoBERTa-base | 125M | 90.45 | - | - |
| (Cao et al., 2023) (zero-shot) | GPT-3.5 | - | - | 35.14 | 33.51 |
| (Cao et al., 2023) (zero-shot) | Llama-2 (7B) | - | - | 31.93 | 28.58 |
| GPT-4 (zero-shot) | GPT-4 | - | 93.91* | 68.69 | 64.17 |
| LoRA (fine-tuned, ours) | Llama-2 (7B) | 5M | 94.29* | 70.17 | 66.59 |
| BiSaGA (w/o feature-compression) | Llama-2 (7B) | 150M | 94.50* | 71.37 | 68.61 |
| BiSaGA | Llama-2 (7B) | 5M | **95.08*** | **73.57** | **71.89** |

Table 3: 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).

The experimental results, as displayed in Table 3, show that our BiSaGA outperforms all other baseline models, including GPT-4, on both the FEVER (English) and CHEF (Chinese) datasets. Specifically, BiSaGA achieves a label accuracy (LA) of 95.08% on FEVER and 73.57% on CHEF, along with an F1 score of 71.89%. 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, which indicates that these two are not incapable of this task.

Compared to the LoRA fine-tuned Llama-2 model, BiSaGA demonstrates a notable improvement, with relative gains of +0.79% and +3.40% in label accuracy (LA) on the FEVER and CHEF datasets, respectively. This underscores that our framework enables better adaptation of Llama-2 to evidence-based fact-checking tasks compared to LoRA, thus proving the effectiveness of our adaptation mechanism.

Compared to the framework without the feature-compression mechanism, BiSaGA shows relative improvements of +0.58% and +2.20% in label accuracy (LA) on FEVER and CHEF, respectively. One possible explanation is that the original adapter without feature-compression struggles in data-scarce scenarios (Zoph et al., 2016; Hedderich et al., 2021), potentially making full-parameter fine-tuning susceptible to undertraining and overfitting (Mahabadi et al., 2021). BiSaGA circumvents these issues through its lightweight design, introducing only 5M parameters compared to the 7B of the base model.

## 4 Analysis

### 4.1 Reversal Curse

We analyze those instances in CHEF-RC that possibly lead to the Reversal Curse as shown in Table 4. Preliminary estimates show that in the SUP and REF classes, errors in GPT-4 caused by the Reversal Curse accounted for 39.64% and 59.46% of total errors, respectively, summing up to 49.55%. In comparison, our framework reduced these errors to 27.03% in the SUP class and 17.65% in the REF class, with a combined total of 24.07%. This result verifies that BiSaGA works great against the Reversal Curse, marking the usefulness of its bidirectional modeling.

| | | CHEF-RC | |
|---|---|---|---|
| | | SUP | REF |
| BiSaGA | SUP | 35 | 3 |
| | REF | 10 | 57 |
| | NEI | 0 | 0 |
| | R (%) | 92.11 | 85.07 |
| | P (%) | 77.78 | 95.00 |
| GPT-4 | SUP | 29 | 7 |
| | REF | 6 | 48 |
| | NEI | 10 | 5 |
| | R (%) | 80.56 | 88.89 |
| | P (%) | 64.44 | 80.00 |

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

| Method | FEVER | CHEF | |
|---|---|---|---|
| | LA (%) | LA (%) | F1 (%) |
| BiSaGA | 95.08 | 73.57 | 71.89 |
| w/o BiSaGA $_1$ | 94.17 | 69.57 | 65.54 |
| w/o BiSaGA $_2$ | 94.36 | 71.27 | 67.75 |
| w/o BiSaGA $_3$ | 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.

### 4.2 Ablation Analysis

In this section, we perform ablation experiments on the proposed hierarchical adaptation structure. The results are presented in Table 5. Our findings indicate that each layer in BiSaGA enhances performance, affirming its effectiveness. The layers are numbered from 1 to 3, from the front to the back of the model. Among all the layers, Layer 1 is the most essential. Removing this layer results in Llama performing even worse than the original LoRA version, highlighting the superiority of our method's attention mechanism with a small sliding window. In FEVER, Layer 2 has a more significant impact on the results than Layer 3, while it is the other way around in CHEF.

### 4.3 Attention Pattern

To gain deeper insights into how the bidirectional sparse graph attention influences the final Value representations, we study its attention pattern, as shown in Figure 2. 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 BiSaGA 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 BiSaGA 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.

### 4.4 Case Study

In this section, we conduct case analyses on random samples of the CHEF-RC dataset to evaluate the practical effectiveness of our framework compared to LoRA. For the English study, we translated these samples to align with Chinese and asked the English fine-tuned models to answer these questions. The results are presented in Table 6 and Table 7. Apart from predicted probabilities, we perform norm calculations to compare the amplification effect of BiSaGA and LoRA, which illustrates how much the features change compared to those in the original model.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 946-959, Jinan, China, August 11-14, 2025.

 

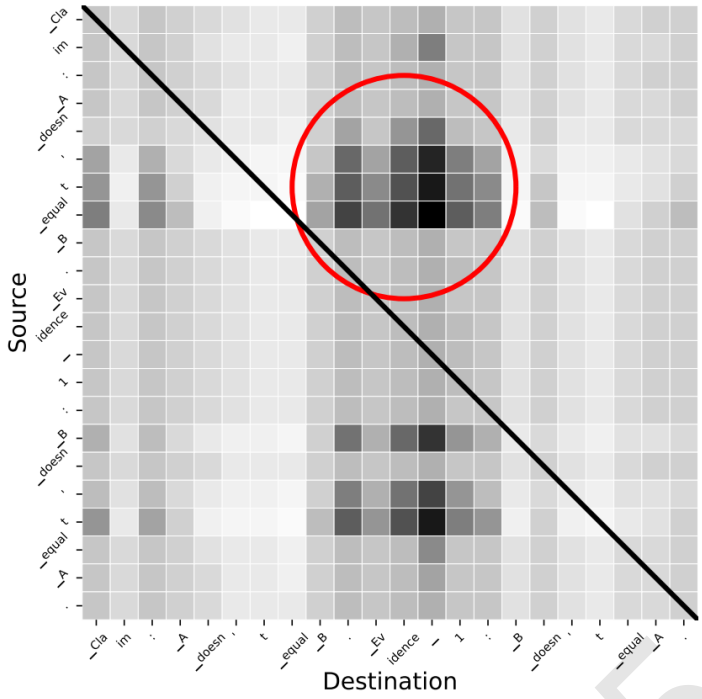


Figure 2: Attention illustration on an instance of our proposed BiSaGA framework.

| ID | Label | GPT-4 Prediction | LoRA $\|\Delta V\|_1$ | LoRA $\|\Delta V\|_2$ | LoRA Prediction | LoRA Probability (%) | BiSaGA $\|\Delta V\|_1$ | BiSaGA $\|\Delta V\|_2$ | BiSaGA Prediction | BiSaGA 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 6: Chinese case study of CHEF-RC.

The results indicate that the amplification effect of our embedded module is only between 1/5 and 1/4 of that in the LoRA module. However, in some cases, like case 686, our BiSaGA achieved the correct prediction, whereas LoRA made an incorrect prediction. Our bidirectional sparse graph attention adapter (BiSaGA) achieves correct predictions with higher probabilities than LoRA in most cases, despite having lower amplification effects. These findings suggest that our adaptation is more compact and effective than LoRA, highlighting the superiority of our framework.

Another interesting finding is that cases predicted as NEI tend to exhibit lower $l_1$ and $l_2$ variations compared to other cases. Additionally, for each class in each model, predicted cases with low probabilities usually have smaller $l_1$ and $l_2$ variations compared to those with high prediction probabilities.

## 5 Related Work

### 5.1 Previous Works on Evidence-based Fact-checking

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 (Gupta and Srikumar, 2021; Hu et al., 2023b). Some approaches tried to solve this task with representations of graph structure. (Zhou et al., 2019)



 

| ID | Label | GPT-4 Prediction | LoRA $\|\Delta V\|_1$ | LoRA $\|\Delta V\|_2$ | LoRA Prediction | LoRA Probability (%) | BiSaGA $\|\Delta V\|_1$ | BiSaGA $\|\Delta V\|_2$ | BiSaGA Prediction | BiSaGA 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 7: English case study of CHEF-RC.

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.

### 5.2 LLM Attempts on Evidence-based Fact-checking

With the advancements of LLMs, numerous attempts have been made at evidence-based fact-checking. (Cao et al., 2023) evaluated the fact verification performance of gpt-3.5-turbo and Llama2-7b. FactLlama (Cheung and Lam, 2023) combined Llama with external evidence retrieval to bridge the gap between model knowledge and up-to-date context. HiSS (Zhang and Gao, 2023) employed a Hierarchical Step-by-Step method with text-davinci-003 to break down claims into sub-claims and verify each via multiple question-answering steps. (Hu et al., 2023a) used Llama-7B and gpt-3.5-turbo to test on the Pinocchio benchmark with 20K factual questions. (Quelle and Bovet, 2023) utilized GPT-3.5 and GPT-4 for fact-checking by querying, retrieving context, and making decisions with cited reasoning. (Choi and Ferrara, 2023) designed a framework for automating the claim-matching phase using various GPT and Llama models on a GPT-4 generated dataset of simulated social media posts.

### 5.3 Integrating Graphs with LLMs

Many studies have attempted to combine LLMs with graph neural networks. (Chen et al., 2023) explored the potential of LLMs in graph neural networks through two pipelines: enhancing node features with LLMs and using LLMs as standalone predictors. (Guo et al., 2023) conducted an empirical study to assess LLMs' comprehension of graph data, using various tasks to evaluate their graph understanding. They introduced a framework combining LLMs and graph-structured data, utilizing graph description language with prompt engineering. Graph of Thoughts (GoT) (Besta et al., 2023) advanced LLM prompting by modeling LLM-generated information as graphs, where thoughts are vertices and dependencies are edges. (He et al., 2023) leveraged LLMs to capture textual information as graph features to enhance GNN performance.

### 5.4 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.

## 6 Conclusions and Future Works

We proposed the BiSaGA framework, a bidirectional sparse graph attention adapter for LLMs. This framework introduces bidirectional attention to hierarchical sparse graphs for enhanced information aggregation and efficient fine-tuning. Our method successfully overcomes the Reversal Curse with bidirectional attention adaptation, achieving superior performance with aggregated information. As a result, we improved model capabilities, surpassed GPT-4, and set new state-of-the-art (SOTA) results in the evidence-based fact-checking task. Our framework can address the Reversal Curse in various reasoning tasks, representing a significant advancement. We are committed to exploring this promising approach in other fields.

## 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 Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, November. 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 Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors, *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 37–48, Dublin, Ireland, May. 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 Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors, *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 58–65, Dubrovnik, Croatia, May. 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 Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June. Association for Computational Linguistics.

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, August. 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, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *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, July. Association for Computational Linguistics.

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, 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.

Yuqing Lan, Zhenghao Liu, Yu Gu, Xiaoyuan Yi, Xiaohua Li, Liner Yang, and Ge Yu. 2024. Multi-evidence based fact verification via a confidential graph neural network.

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 Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. 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 Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online, July. Association for Computational Linguistics.

Weiyao Luo, Junfeng Ran, Zailong Tian, Sujian Li, and Zhifang Sui. 2024. FaGANet: An evidence-based fact-checking model with integrated encoder leveraging contextual information. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, May. ELRA and ICCL.

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.

OpenAI. 2023. Gpt-4 technical report.

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 James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal, editors, *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 132–137, Brussels, Belgium, November. 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 Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online, July. 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 Marilyn Walker, Heng Ji, and Amanda Stent, editors, *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, June. 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 Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2596–2601, Lisbon, Portugal, September. 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.

Amelie Wuehrl, Lara Grimminger, and Roman Klinger. 2023. An entity-based claim extraction pipeline for real-world biomedical fact-checking. In Mubashara Akhtar, Rami Aly, Christos Christodoulopoulos, Oana Cocarascu, Zhijiang Guo, Arpit Mittal, Michael Schlichtkrull, James Thorne, and Andreas Vlachos, editors, *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 29–37, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Shuo Yang, Bardh Prenkaj, and Gjergji Kasneci. 2024. Razor: Sharpening knowledge by cutting bias with unsupervised text rewriting.

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A two-wing optimization strategy for evidential claim verification. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium, October-November. 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.

Liwen Zheng, Chaozhuo Li, Haoran Jia, and Xi Zhang. 2025. Reasoning paths as signals: Augmenting multi-hop fact verification through structural reasoning progression.

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 Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy, July. Association for Computational Linguistics.

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