Retrieve, Rethink, and Review: Cross-Granularity Retrieval for Fact Verification

Anonymous ACL submission

Abstract

The dissemination of misinformation on online platforms has necessitated the development of automatic fact verification systems. Recent studies leverage semantic features of both claims and evidence to make predictions. However, these methods hypothesize that evidence is always available and accessible, which is undoubtedly impossible in real-world circumstances. Recent studies attempt to use retrieval-augmented approaches to retrieve relevant evidence to conduct fact verification tasks. However, these methods typically use the entire statement as a query to retrieve evidence, which may lead to missing relevant results. Besides, some studies utilize decomposed claims as queries, but they omit the filtering process, which may retrieve redundant information. Thus, to solve these challenges, we propose a novel Cross-granularity Retrieval-Augmented Network (CRAN) for open-domain fact verification. Specifically, we first utilize an LLM-based decomposer to divide the claims into atomic facts, facilitating sufficient retrieval. Besides, we leverage a novel reranking method to filter the redundant evidence. Then, we design a bipartite graph to fuse claim-evidence representations and make predictions. The experimental results on four common-used datasets demonstrate the effectiveness and superiority of our model.

1 Introduction

The dissemination of misinformation and disinformation has become a critical problem in the age of information overload, where false or misleading claims can easily spread across various platforms. This stimulates the development of automatic fact verification methods, aiming to verify the truthfulness of a given claim as a safeguard to protect public knowledge and trust in information (Chen et al., 2025; Kanaani, 2024; Bazaga et al., 2024).

Traditional methods of fact verification typically involve automated systems that assess the veracity of a claim using authorized or fact-checked evidence (Guo et al., 2022; Zeng et al., 2021). These methods hypothesize that the gold evidence of each claim is available and accessible, and they cannot deal with the data without evidence. However, in real-world circumstances, the relevant evidence needs to be retrieved dynamically from open sources or databases like Wikipedia[1] and PolitiFact[2]. Hence, traditional approaches are hard to apply to real-world fact verification systems.

Therefore, retrieval-augmented fact verification methods have been proposed to tackle this problem (Fan et al., 2024; Chakrabarty et al., 2018; Shi et al., 2021). These methods utilize the claim as a query to retrieve relevant evidence from knowledge bases or open sources and verify the truthfulness of the claim with the retrieved information. These methods are more capable of dealing with real-world scenarios. However, they tend to ignore that one claim may contain multiple facts to be verified by using the whole claim as one query. For example, the statement *Shands Hospital of Gainesville Florida has confirmed its first case of coronavirus* has 2 facts to be checked. Specifically, *Shands Hospital located in Gainesville Florida* and *Shands Hospital confirmed the first case of coronavirus* are 2 hidden atomic statements whose truthfulness is unclear.

Taking this circumstance into account, some methods divide claims into several atomic propositions to help retrievers find more sufficient and comprehensive evidence (Zhang and Gao, 2024, 2023). They utilize decoder-based models like T5 (Raffel et al., 2019) to extract atomic propositions. Recent studies leverage Large Language Models (LLMs) as atomic proposition extractors. However, divided propositions may contain the same fact points and these methods overlook the importance

[1] https://www.wikipedia.org/
[2] https://www.politifact.com/

of filtering redundant evidence. Thus the retriever will find redundant evidence. This redundancy may divert the model's attention, thereby causing other more critical evidence to be overlooked. This dispersion of focus could potentially lead to suboptimal weighting of information priorities during the analytical process.

To solve the aforementioned problems, we propose a novel Cross-granularity Retrieval-Augmented Network (CRAN) for open-domain fact verification. Specifically, we decompose the claim into several propositions through an LLM-based decomposer to retrieve sufficient and comprehensive information. Different from existing methods, our decomposition process directly obtains interrogative atomic propositions instead of declarative atomic propositions. This is because interrogative sentences are easier to generate queries without plenty of semantic modifications. Besides, we leverage a novel rethinking method to filter redundant evidence, taking 3 dimensions into account, relativity, utility and redundancy. This avoids the negative impact of information fusion and the model's attention weight. Then, we design a claim-evidence bipartite graph to learn fused representations and make predictions to verify truthfulness. Different from the traditional bipartite graph, we construct two subgraphs using the claim set and the evidence set respectively. This modification can formulate the hierarchical information and facilitate information integration. We use both evidence retrieval and fact verification datasets to investigate the effectiveness of CRAN. The experimental results demonstrate the CRAN outperforms other state-of-the-art methods and is capable of solving open-domain fact verification tasks.

Our major contributions are as follows:

- We propose a novel Cross-granularity Retrieval-Augmented Network for open-domain fact verification tasks. We use the decomposition method and reordering process to retrieve comprehensive information to predict the verdict, as well as to eliminate the impact of redundancy.

- We design a new Bipartite graph integration module to fuse claim and evidence information and learn comprehensive representations. This module enables the model to better understand the relations between propositions and evidence.

- To evaluate the performance of the proposed method, we carry out experiments on 4 datasets. Our model outperforms the comparison methods, which demonstrates the effectiveness and superiority of the proposed model.

2 Related work

Fact verification aims to predict the verdicts of check-worthy claims with several retrieved evidence. Traditional fact verification approaches only utilize textual information to make predictions (), which fails to deal with claims that need multi-hop consideration. Hence, multi-hop fact verification has become a research hotspot. Besides, LLMs have made significant developments and have been applied to fact verification tasks. In this section, we will report on the related work in these three research fields.

2.1 Fact verification

Research on unimodal fact verification typically involves verifying text-only claims using textual evidence, such as metadata of the claim, documents retrieved from knowledge bases, or tabular evidence (Wang, 2017; Aly et al., 2021; Panchendrarajan and Zubiaga, 2024; Gong et al., 2024).

Wang (2017) incorporated additional metadata as external evidence, such as the speaker's profile, locations and communities, to verify claims using a Convolutional Neural Network. By harnessing entities derived from textual materials, Chen et al. (2021) built entity graphs with the aim of acquiring more detailed data representations. Meanwhile, certain scholars have endeavoured to exploit structured sources, including tabular information, to achieve enhanced results. For example, Gu et al. (2022) serialized table evidence to convert table evidence into a sequential form and merged it with the claim to determine its validity. Wang et al. (2021) learnt the salient semantic representations for fact verification to deal with the unbalanced vocabulary of statements and evidence.

These approaches leverage various claim-evidence interaction methods to deal with text-only fact verification and demonstrate satisfactory performance on unimodal fact verification.

2.2 Retrieval-augmented fact verification

Studies on retrieval-augmented fact verification tend to incorporate effective retrieval methods to

obtain relevant evidence and solve the open-domain fact verification tasks (Chakrabarty et al., 2018; Shi et al., 2021; Chen et al., 2022).

FEVER (Chakrabarty et al., 2018) first incorporated the retrieval methods with verification tasks and built the FEVER dataset for further investigation. Since then, retrieval-augmented fact verification has become a hotspot. Shi et al. (2021) tackled table-based fact verification by establishing an evidence retrieval and reasoning framework. This approach effectively integrates evidence retrieval with logical reasoning to enhance the accuracy of fact verification. Chen et al. (2022) proposed GERE that improves fact verification by generating evidence and its title more efficiently and dynamically, addressing the shortcomings of traditional retrieval methods. RAV (Zheng et al., 2024) added a simple similarity-based reordering process after the coarse-grained retrieval, avoiding irrelevant or redundant evidence to impact the performance. RAFTS (Yue et al., 2024) utilizing LLMs to retrieve document-level evidence and combine sparse and dense retrieval methods to enhance the retrieval performance.

These approaches leverage various retrieval methods to enhance the model's capability of searching and extracting relevant information as evidence and to solve open-domain fact verification.

3 Methodology

This section presents the Cross-granularity Retrieval-Augmented Network (CRAN) in detail for open-domain fact verification. We begin by defining the task and introducing the overall framework of CRAN. Then, we'll go over the details of the proposed method.

3.1 Task definition

Open-domain fact verification aims to retrieve relevant information from open knowledge bases as evidence and verify the truthfulness of claims. Let $\mathcal{C} = \{C_1, C_2, \cdots, C_{|C|}\}$ and KB be the claim set consisting of claims that need to be verified and the open knowledge base. Each C_i has a truthfulness label $y \in \mathcal{Y}$. The task first requires to retrieve m relevant documents $\mathcal{D}_i = \{D_i^1, D_i^2, \cdots, D_i^m\}$ as the evidence set for each claim C_i. Then it needs to find a function $\mathcal{F} : (\mathcal{C}, \mathcal{D}) \rightarrow \mathcal{Y}$ that maps the data to the label set and makes predictions.

3.2 Overall architecture

We aim to decompose the claim into atomic propositions and retrieve relevant evidence to verify the claim's truthfulness. Hence, we propose a novel *Cross-granularity Retrieval-Augmented Network* for open-domain fact verification. Fig. 1 illustrates the overall architecture of CRAN, which consists of the following modules:

- **LLM-based claim decomposition**: We fine-tune an LLM to decompose the claim into atomic propositions for fine-grained evidence retrieval to obtain sufficient and comprehensive evidence.
- R^2 **Document-level retrieval**: We leverage a retrieval model to retrieve relevant information from an open knowledge base, and use a rethinking approach to filter redundant noises.
- **Bipartite graph integration**: We construct a subgraph-based bipartite graph using claims and evidence to fuse information and obtain comprehensive graph representations.
- **Fact classification**: We utilize an MLP-based classifier to verify the truthfulness with the graph representation of each claim.

3.3 LLM-based claim decomposition

Claim decomposition is proved to be the better way to retrieve evidence as comprehensively as possible, and LLMs demonstrate the great capability of dealing with this task. Hence, we fine-tune Llama3.1-8b (Grattafiori et al., 2024) to fully explore the capability of LLM to decompose the claim correctly and sufficiently, instead of directly utilizing the knowledge of LLM learned in pre-training. Specifically, instead of utilizing declarative atomic propositions, we train the model to generate interrogative propositions directly, which is more suitable to be a query. To this end, following Hu et al. (2021), we deliberately choose the decomposition-based fact verification dataset AVeriTeC (Schlichtkrull et al., 2023) and utilize the Low-Rank Adaptation (LoRA) mechanism to fine-tune the LLM to obtain atomic propositions[3]. Consequently, for each C_i, we obtain its proposition set $\mathcal{P}_i = \{P_i^1, P_i^2, \cdots, P_i^k\}$ for the document-level retrieval.

[3]The reason we utilize LoRA rather than SFT to fine-tune Llama3.1 is that the scale of the training data is relatively low.

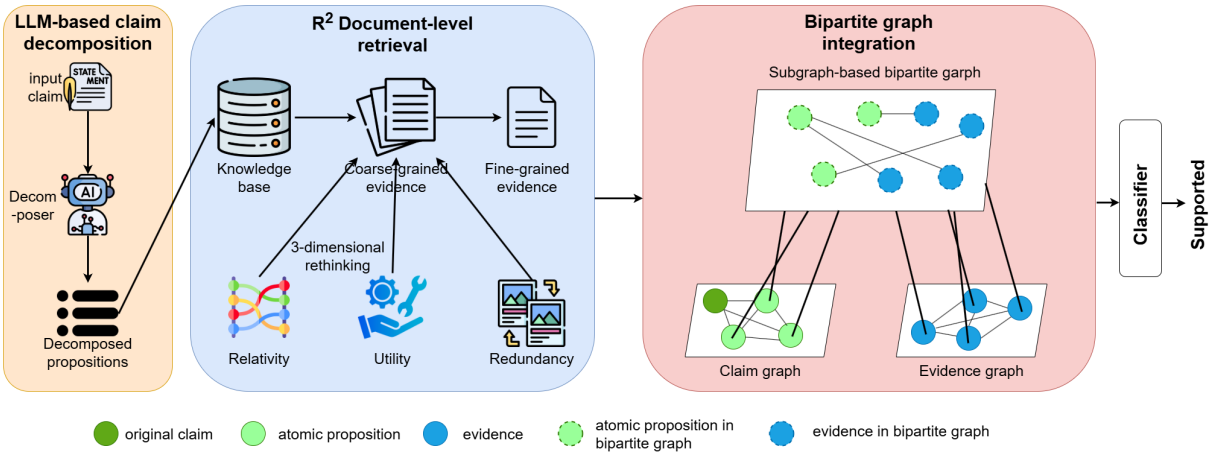

Figure 1: The framework of CRAN. The model takes only the claim as input. It can be divided into four modules: (1) LLM-based claim decomposition for claim decomposition, (2) R^2 Document-level retrieval for evidence retrieval, (3) Bipartite graph integration for information integration, and (4) Fact classification for truthfulness prediction.

3.4 R^2 Document-level retrieval

For document-level retrieval, we design a 2-stage retrieval method. First, we utilize a retrieval model to retrieve relevant documents as comprehensively as possible. Then we leverage a rethinking method to filter the duplicates to eliminate the impact of redundancy.

Document-level retrieval In this stage, our objective is to retrieve more relevant information for further verification. Hence, we use a retrieval model E5-base (Wang et al., 2022) as the document-level retriever to extract relevant information. Specifically, given the claim C_i and its propositions $\mathcal{P}_i = \{P_i^1, P_i^2, \cdots, P_i^k\}$, we set each P_i as a query and retrieve the top-n relevant document to obtain the coarse-grained evidence set $\hat{\mathcal{D}}_i = \{\hat{D}_i^1, \hat{D}_i^2, \cdots, \hat{D}_i^n\}$.

Document-level rethinking As aforementioned in section 1, the redundancy of evidence will have a negative impact on model attention arrangement. Thus, we utilize a rethinking approach to filter the duplicates and eliminate this problem.

In detail, we set a score s_{ij}^{re} using the score obtained during the retrieval stage for each retrieved document $\hat{D}_i^j$ as the retrieval score. Then, we use an encoder to obtain the embeddings c_i and $\hat{d}_i^j$ of the claim C_i and each $\hat{D}_i^j$, and obtain the verification score s_{ij}^{ve} by the following equation:

$$s_{ij}^{ve} = cos(c_i, \hat{d}_i^j). \quad (1)$$

cos denotes the function to calculate cosine similarity. This score reflects the importance of each document to the original claim during the verification task. Besides, we design a redundancy score s_{ij}^{rd} to represent the redundancy extent of each $\hat{D}_i^j$ in the coarse-grained evidence set. This score is calculated by:

$$s_{ij}^{rd} = cos(\hat{d}_i^j, \hat{d}_i). \quad (2)$$

$\hat{d}_i$ denotes the embedding matrix of all relevant document to proposition i. This score quantitatively measures the semantic similarity between documents, facilitating to filter the duplicate noises.

With these scores, we can reorder the documents and select top-m more relevant and more contributive documents for fact verification. We first integrate them to obtain the final score s_{ij} by:

$$s_{ij} = \mathbf{A}[s_{ij}^{re} || s_{ij}^{ve} || (1 - s_{ij}^{rd})]^T. \quad (3)$$

$\mathbf{A} = [a_1, a_2, a_3]$ is the hyperparameter that denotes the weight for each score. This weighted fused score takes three dimensions into account, not only to ensure the effectiveness and helpfulness of the retrieved documents to the verification task but also to eliminate the redundancy in the final fine-grained evidence set $\mathcal{D}_i = \{D_i^1, D_i^2, \cdots, D_i^m\}$.

3.5 Bipartite graph integration

After retrieving relevant evidence, we then design a novel graph-based fusion method to integrate information for fact verification. This module can be divided into 2 processes. In the first process, we construct a subgraph-based bipartite graph to structure the information between claim, proposition and evidence. Then we leverage a modified Graph Attention Network (m-GAT) model to learn comprehensive graph representations.

Subgraph-based bipartite construction For each claim C_i, we merge the claim and its relevant propositions into a set $\hat{\mathcal{P}}_i = \{C_i, P_i^1, P_i^2, \cdots, P_i^k\}$. Then, we utilize $\hat{\mathcal{P}}_i$ and its relevant evidence set $\mathcal{D}_i$ as nodes to construct the claim graph and the evidence graph respectively. These two graphs are all fully connected for abundant message passing. Upon these graphs, we construct a bipartite graph BG_i to bridge the relations between propositions and relevant documents. Specifically, **the node set** consists of the proposition set $\mathcal{P}_i$ without the original claim and the fine-grained evidence set $\mathcal{D}_i$. Because we directly use the proposition as the query to retrieve evidence, we hypothesize that the relation between the proposition and evidence is closer to that between the claim and evidence. Thus, **the edge between the proposition and evidence** exists only if the evidence is retrieved by this proposition. Finally, we obtain the subgraph-based bipartite $BG_i = \{V_i, E_i\}$, where $V_i = \{\mathcal{P}_i, \mathcal{D}_i\}$ and $E_i = \{e = 1 | \text{if evidence is retrieved by proposition.}\}$.

Modified graph attention fusion Given the bipartite graph BG_i, it is hard to directly leverage existing graph fusion methods to learn the representations. Hence, we propose a novel modified Graph Attention Network (m-GAT) to fuse information and learn comprehensive representations.

To initialize the graph, we utilize the textual encoder to obtain the embeddings c_i for the claim, p_i^j for each proposition, and d_i^q for evidence respectively. For the l-th layer integration, we first integrate the information of the evidence graph and mask other edges by:

$$d_i^{q(l)} = \alpha_{q,q}\Theta d_i^{q(l-1)} + \sum_{j \in \mathcal{N}(q)} \alpha_q \Theta d_i^{j(l-1)}. \quad (4)$$

$d_i^{q(l)}$ and $d_i^{q(l-1)}$ denote the evidence representation of the l-th and $(l-1)$-th layer respectively. Θ is a learnable matrix. $\mathcal{N}(q)$ denotes the neighborhood set of d_i^q. α denotes the attention weight calculated by:

$$\alpha_{q,j} = \frac{\exp(a^T \sigma(\Theta[d_i^q || d_i^j]))}{\sum_{k \in \mathcal{N}(q) \cup \{q\}} \exp(a^T \sigma(\Theta[d_i^q || d_i^k]))}. \quad (5)$$

a^T is a learnable vector.

Then, we mean to update the bipartite node representations by:

$$h_i^{j(l)} = \beta_{j,j}\Theta h_i^{j(l-1)} + \sum_{k \in \mathcal{N}(j)} \beta_k \Theta h_i^{k(l-1)}. \quad (6)$$

$h_i^{j(l)}$ denotes the l-th representation of the bipartite graph, including proposition nodes $p_i^j(l-1)$ and evidence nodes $d_i^{q(l)}$. β denotes the attention weight which is calculated by eq. 5. It is important that we use the updated evidence representations as the input during the fusion of the bipartite graph because the relation between evidence is useful for understanding the relation between propositions and contributes to verdict prediction.

Finally, we integrate the information of the claim graph by eq. 4 and eq. 5. After this, the l-th layer integration is complete and we utilize the l-th representation of the claim $c_i^{(l)}$ as the l-th layer graph representation.

3.6 Fact classification

To predict the label of the given claim, we utilize the final layer graph representation $c_i^{(L)}$ as the fused representation that contains all information to predict the truthfulness. The prediction process is carried out as follows:

$$\hat{y} = softmax(W^1 \sigma(W^0) c_i^{(L)}). \quad (7)$$

L denotes the last layer of the graph fusion module. W^0 and W^1 are trainable parameters. $\hat{y}$ is the predicted label. We train our model by minimizing cross-entropy loss to learn the prediction of the categories.

4 Experiment

4.1 Datasets and evaluation metrics

Datasets To evaluate the effectiveness of our proposed CRAN for both evidence retrieval task and fact verification task, we choose MS MACRO (Nguyen et al., 2016) and Check-COVID (Wang et al., 2023) datasets as the retrieval datasets, and LIAR (Wang, 2017) and FEVER (Chakrabarty et al., 2018) datasets as the verification datasets. MS MACRO is a large-scale QA dataset that requires the first retrieval of relevant documents or passages to find answers. Check-COVID is a fact-checking dataset that contains claims about COVID-19 and need to retrieve evidence to fact-check them. LIAR and FEVER are two fact-checking datasets that do not offer the evidence directly, which need to first retrieve relevant information and then predict the verdict of each claim.

Evaluation metrix For the evidence retrieval task, the adopted metrics are NDCG and Recall (i.e., N@k and R@k) with $k \in [3, 5]$. For the fact

Model	MS MACRO				Check-COVID			
	N@3	R@3	N@5	R@5	N@3	R@3	N@5	R@5
TF-IDF	0.531	0.613	0.562	0.687	0.363	0.427	0.385	0.480
BM25	0.746	0.801	0.760	0.836	0.395	0.467	0.426	0.545
DPR	0.793	0.850	0.797	0.903	0.411	0.477	0.457	0.588
E5-base	0.855	0.895	0.865	0.920	0.584	0.679	0.609	0.741
RAFTS	0.858	0.896	0.868	0.920	0.631	0.712	0.646	0.750
RAV	0.856	0.896	0.867	0.921	0.630	0.709	0.640	0.738
CRAN (Ours)	**0.863**	**0.905**	**0.878**	**0.926**	**0.658**	**0.730**	**0.672**	**0.775**

Table 1: Results of the evidence retrieval task. **Bold** denotes the best performance and Underline denotes the second-best performance.

verification tasks, we use Accuracy and Macro F1 score as the evaluation metrics.

4.2 Baselines

We compare our CRAN to several existing methods in evidence retrieval and fact verification tasks.

Evidence retrieval task We choose 2 sparse retrieval methods TF-IDF and BM25 and 4 dense retrieval methods DPR (Karpukhin et al., 2020), E5-base (Wang et al., 2022), RAFTS (Yue et al., 2024), and RAV (Zheng et al., 2024) to evaluate the effectiveness of our proposed R^2 document-level retrieval method. DPR is the first retriever only leverage dense embedding to retrieve documents. E5-base is an LLM-based retriever. RAFTS leverages a 2-stage retrieval method that combines sparse retriever and dense retriever. RAV incorporates reordering after obtaining the coarse-grained retrieved documents.

Fact verification task We select 2 transformer-based methods BERT (Devlin et al., 2019) and DeBERTa (He et al., 2021), and 3 graph-based methods GEAR (Zhou et al., 2019), KGAT (Liu et al., 2020), and SAGP (Si et al., 2023) to evaluate the usefulness of our proposed Bipartite graph integration. GEAR leverages a graph neural network to fuse claim and evidence features and make predictions. KGAT utilize a kernel-based graph fusion method to predict the truthfulness of claims. SAGP leverages a perturbed graph neural network and selects rational subgraphs to make predictions and give explanations.

4.3 Implementation details

We use a Tesla V100-PCIE GPU with 32GB memory for all experiments and implement our model via the Pytorch framework[4]. The number of attention heads is set to 4, and the number of GAT convolutional layers is 3. The number of decomposed propositions is up to 5. The number of retrieved coarse-grained documents n is set to 20 and the number of fine-grained documents m is set to 5. The batch size is 8. We set the learning rate as 2e-5. We employ E5-base (Wang et al., 2022) as the textual embedding model.

Model	FEVER		LIAR	
	Acc	F1	Acc	F1
BERT	73.67	72.80	77.49	77.18
DeBERTa	73.90	73.19	78.03	78.01
GEAR	74.84	74.79	80.96	80.41
KGAT	78.29	77.11	83.97	83.62
SAGP	82.10	82.15	85.50	85.48
CRAN (Ours)	**84.45**	**84.37**	**86.73**	**86.67**

Table 2: Results of the fact verification task. Acc and F1 denote the Accuracy and MACRO F1 score respectively. **Bold** denotes the best performance and Underline denotes the second-best performance.

5 Result

5.1 Overall performance

We conduct the experiments on 4 datasets on evidence retrieval and fact verification tasks. The experimental results are shown in Table 1 and Table 2.

Evidence retrieval Table 1 demonstrates the experimental results of the evidence retrieval task. It can be observed that CRAN outperforms other retrieval models. Specifically, compared to sparse retrieval models like TF-IDF and BM25, all dense retrieval methods perform better and there is a huge

[4]Because of the hardware limitation, we decompose the claim in the preprocessing so that we do not load the fine-tuned LLM in GPU during training and testing.

Model	MS MACRO				Check-COVID			
	N@3	R@3	N@5	R@5	N@3	R@3	N@5	R@5
CRAN	**0.863**	**0.905**	**0.878**	**0.926**	**0.658**	**0.730**	**0.672**	**0.775**
-w/o rethinking	0.852	0.892	0.863	0.915	0.626	0.702	0.637	0.732
-w/o relativity	0.859	0.898	0.871	0.923	0.649	0.718	0.666	0.763
-w/o utility	0.858	0.901	0.872	0.923	0.650	0.720	0.667	0.766
-w/o redundancy	0.854	0.893	0.865	0.918	0.630	0.705	0.640	0.733

Table 3: Results of the ablation study of evidence retrieval task. **Bold** denotes the best performance.

Model	FEVER		LIAR	
	Acc	F1	Acc	F1
CRAN	**84.45**	**84.37**	**86.73**	**86.67**
-w/o claim graph	83.90	83.97	86.59	86.55
-w/o evidence graph	83.78	83.79	86.48	86.40
-w/o bipartite graph	83.45	83.34	86.41	86.38

Table 4: Results of the ablation study of fact verification task. Acc and F1 denote the Accuracy and MACRO F1 score respectively. **Bold** denotes the best performance.

gap between sparse retrieval models and dense retrieval methods. It indicates that dense retrieval methods are more capable of the evidence retrieval task. Besides, compared to the RAFTAS and RAV which either use the combination of sparse and dense retrievers or utilize a reordering process without decomposition, CRAN has a better performance, elucidating that the balanced combination of decomposition and rethinking is beneficial to the evidence retrieval task.

Fact verification Table 2 shows the experimental results of the fact verification task. We utilize the same retrieved evidence as the input evidence to conduct the experiments. Our proposed CRAN outperforms other methods significantly. On the FEVER dataset, CRAN improves both accuracy and F1 value by 2.35% and 2.22% respectively. On the LIAR dataset, CRAN also improves accuracy and F1 value by 1.23% and 1.19% respectively. This indicates the effectiveness of CRAN in fact verification tasks. Besides, compared to transformer-based methods, graph-based methods are more capable of understanding the inner relations between claims and evidence.

Overall, the experimental results of evidence retrieval and fact verification tasks demonstrate that our model CRAN is able to deal with these two tasks and outperforms the state-of-the-art methods, elucidating the effectiveness and superiority of CRAN.

5.2 Ablation study

We also conduct the ablation experiments to further investigate the effectiveness of our proposed modules. The experimental results are shown in Table 3 and Table 4.

Table 3 demonstrates the results of evidence retrieval tasks. It can be observed that by removing the rethinking process, the performance of CRAN drops, even lower than baseline models. To further analyze this phenomenon, we remove the 3 dimensions respectively to find out the most decisive dimension. According to the experimental results, the model's performance degrades the most when we remove the redundancy score. It indicates that among these dimensions, redundancy is the most harmful to the retriever.

Besides, we investigate the effectiveness of the subgraph-based bipartite graph. Consistently, removing each graph the performance degrades, elucidating that the relations and information between claims, propositions and evidence are significant in open-domain fact verification tasks. Surprisingly, we hypothesize that claim-proposition relations are more important, but the results show that relations between evidence are more essential. This phenomenon may be attributed to the fact that atomic propositions are derived from the original claims and are merely connected by simple additive relationships. In contrast, the evidence is retrieved based on different atomic propositions, and the early fusion of evidence provides greater benefits for subsequent information integration.

5.3 Module analysis

To further analyze the usefulness of our proposed retrieval method and fusion method, we conduct 2 experiments to illustrate their effectiveness. The experimental results are shown in Fig.2 and Fig.3.

Analysis of retrieval model Fig.2 demonstrates the performance of different retrieval models. *CRAN+RAV* denotes that this model uses the re-

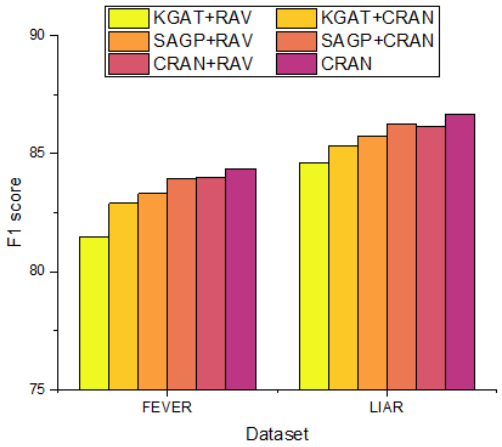

Figure 2: Results of different retrieval methods.*BERT+CRAN* denotes that the retrieval method is CRAN and the fusion method is BERT.

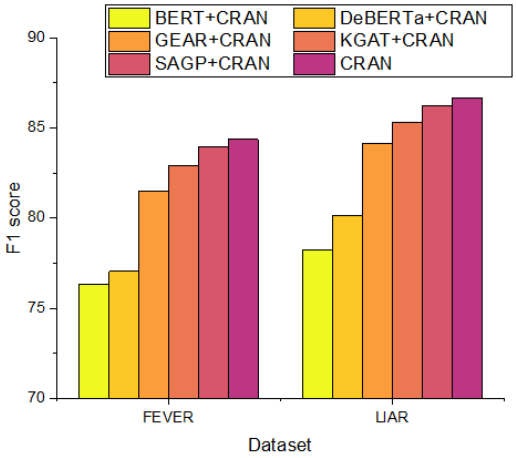

Figure 3: Results of different fusion methods. *BERT+CRAN* denotes that the retrieval model is CRAN and the fusion model is BERT.

trieval method of RAV and the fusion method of CRAN. It is obvious that once the fusion method is fixed, our model outperforms the RAV method on both FEVER and LIAR. This indicates that the CRAN retrieval method is more capable of performing open-domain fact verification tasks. Compared to RAV, it may be because we first decompose the original claim, facilitating the retrieval model to comprehensively extract more relevant information from the knowledge base. Besides, RAV only utilizes similarity as the reordering score, overlooking the other important dimensions when filtering redundant evidence. Significantly, the performance of transformer-based methods still has a huge gap compared to graph-based methods, while the retrieval model of CRAN improves their performances according to Table 2, so do the graph-based methods. This investigation further demonstrates that our proposed R^2 Document-level retrieval can retrieve more crucial evidence and is beneficial to open-domain fact verification tasks, even if the classification method doesn't have satisfactory performance individually.

Analysis of fusion model Fig.3 demonstrates the performance of different fusion models with the retrieval model of CRAN. CRAN outperforms other combined methods. Consistently, the performance of transformer-based methods still has a huge gap compared to graph-based methods. Compared to other graph-based methods, CRAN leverages the advantage of a subgraph-based bipartite graph, that is the precise relations match the propositions and evidence and the hierarchical fusion process that can capture more crucial information.

Overall, these two experiments further demonstrate that the retrieval and fusion methods of CRAN are effective enough to solve open-domain fact verification tasks.

6 Conclusion

In this work, we propose a Cross-granularity Retrieval-Augmented Network to first retrieve the relevant but concise evidence and then leverage a bipartite graph to learn integrated representations for open-domain fact verification. Besides, we design a novel R^2 Document-level retrieval method to retrieve and reorder the information extracted from knowledge bases. Moreover, we design a novel Bipartite graph integration method, constructing a subgraph-based bipartite graph to capture hierarchical information and learn comprehensive representations. The experimental results show that CRAN can effectively retrieve relevant information and make correct predictions for open-domain fact verification tasks. These results highlight the effectiveness and superiority of our proposed model. For future work, we will focus on exploring how we can reduce the retrieval time and systematically combine the retrieval process into the classification.

Limitations

Initially, there is potential for our model to improve the efficiency. As we design the framework, we focus more on accuracy instead of efficiency. How to cut down the time cost of the decomposition and retrieval process is still under investigation. Additionally, in this paper, we regard evidence retrieval and fact verification as two separate tasks, rather

than combining them systematically and adopting dynamic retrieval. Checking while Searching is a hotspot research field in LLMs and RAGs, which may be able to be applied to open-domain fact verification tasks.

References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, and et al. 2021. Feverous: Fact extraction and verification over unstructured and structured information. In *NeurIPS Datasets and Benchmarks 2021*.

Adrián Bazaga, Pietro Lio, and Gos Micklem. 2024. Unsupervised pretraining for fact verification by language model distillation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Tuhin Chakrabarty, Tariq Alhindi, and Smaranda Muresan. 2018. Robust document retrieval and individual evidence modeling for fact extraction and verification. In *Proceedings of the First Workshop on Fact Extraction and VERification, FEVER@EMNLP 2018, Brussels, Belgium, November 1, 2018*.

Chonghao Chen, Fei Cai, Xuejun Hu, and et al. 2021. An entity-graph based reasoning method for fact verification. *Information Processing & Management*, 58(3):102472.

Daoqi Chen, Yaxin Li, Zizhong Zhu, and et al. 2025. Faithful inference chains extraction for fact verification over multi-view heterogeneous graph with causal intervention. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 4634–4645.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, and et al. 2022. GERE: generative evidence retrieval for fact verification. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2184–2189.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and et al. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yue Fan, Hu Zhang, Ru Li, and et al. 2024. FRVA: fact-retrieval and verification augmented entailment tree generation for explainable question answering. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 9111–9128.

Haisong Gong, Weizhi Xu, Shu Wu, and et al. 2024. Heterogeneous graph reasoning for fact checking over texts and tables. In *AAAI*, pages 100–108.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zihui Gu, Ruixue Fan, Xiaoman Zhao, and et al. 2022. Opentfv: An open domain table-based fact verification system. In *SIGMOD*, page 2405–2408.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and et al. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*.

Edward J. Hu, Yelong Shen, Phillip Wallis, and et al. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Mohammadamin Kanaani. 2024. Triple-r: Automatic reasoning for fact verification using language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 16831–16840.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, and et al. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and et al. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7342–7351.

Tri Nguyen, Mir Rosenberg, Xia Song, and et al. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Nat. Lang. Process. J.*, 7:100066.

Colin Raffel, Noam Shazeer, Adam Roberts, and et al. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In

Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Qi Shi, Yu Zhang, Qingyu Yin, and et al. 2021. Logic-level evidence retrieval and graph-based verification network for table-based fact verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 175–184.

Jiasheng Si, Yingjie Zhu, and Deyu Zhou. 2023. Exploring faithful rationale for multi-hop fact verification via salience-aware graph learning. In *AAAI*, pages 13573–13581.

Fei Wang, Kexuan Sun, Jay Pujara, and et al. 2021. Table-based fact verification with salience-aware learning. In *Findings of EMNLP*, pages 4025–4036.

Gengyu Wang, Kate Harwood, Lawrence Chillrud, and et al. 2023. Check-covid: Fact-checking COVID-19 news claims with scientific evidence. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*, page 422–426.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, and et al. 2024. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10331–10343.

Xia Zeng, Amani S. Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Lang. Linguistics Compass*, 15(10).

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 996–1011.

Xuan Zhang and Wei Gao. 2024. Reinforcement retrieval leveraging fine-grained feedback for fact checking news claims with black-box LLM. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 13861–13873.

Liwen Zheng, Chaozhuo Li, Xi Zhang, and et al. 2024. Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 9274–9281.

Jie Zhou, Xu Han, Cheng Yang, and et al. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *ACL*, pages 892–901.