LI^4 : Label-Infused Iterative Information Interacting based Fact Verification in Question-answering Dialogue

Xiaocheng Zhang[1], **Chang Wang**[2], **Guoping Zhao**[3], **Xiaohong Su**[1*]
[1]Faculty of Computing, Harbin Institute of Technology, Harbin, China
[2]School of Computer Science Engineering, South China University of Technology, Guangzhou, China
[3]Zhengzhou Esunny Information Technology Co., Ltd., Zhengzhou, China
22s136029@stu.hit.edu.cn, cswangchang@scut.edu.cn, zhaoguoping@esunny.cc, sxh@hit.edu.cn

Abstract

Fact verification constitutes a pivotal application in the effort to combat the dissemination of disinformation, a concern that has recently garnered considerable attention. However, previous studies in the field of fact verification, particularly those focused on question-answering dialogue, have exhibited limitations, such as failing to fully exploit the potential of question structures and ignoring relevant label information during the verification process. In this paper, we introduce **L**abel-**I**nfused **I**terative **I**nformation **I**nteracting (LI^4), a novel approach designed for the task of question-answering dialogue based fact verification. LI^4 consists of two meticulously designed components, namely the Iterative Information Refining and Filtering Module (IIRF) and the Fact Label Embedding Module (FLEM). The IIRF uses the Interactive Gating Mechanism to iteratively filter out the noise of question and evidence, concurrently refining the claim information. The FLEM is conceived to strengthen the understanding ability of the model towards labels by injecting label knowledge. We evaluate the performance of the proposed LI^4 on HEALTHVER, FAVIQ, and COLLOQUIAL. The experimental results confirm that our LI^4 model attains remarkable progress, manifesting as a new state-of-the-art performance.

Keywords: fact verification, question-answering dialogue, iterative information interacting, label knowledge

1. Introduction

The rampant spread of fake news and rumors can cause mass panic, social unrest, and even war, such as the COVID-19 pandemic(van Der Linden et al., 2020), which constitutes a gravely significant societal challenge. Consequently, the need for fact verification has become increasingly prominent in Natural Language Processing (NLP). Given a claim, the goal of fact verification is to judge the veracity of the claim based on a series of evidence (Guo et al., 2022; Glockner et al., 2022). Most of the existing studies mainly focus on sources like news articles, structured tables(Zhou et al., 2022; Gu et al., 2022), and Wikipedia passages(Rashkin et al., 2017; Bekoulis et al., 2021), while rarely considering the fact verification in the question-answering dialogue. However, as social media platforms(e.g. Twitter, Weibo, and TikTok) gradually become one of the main ways to publish and obtain information, the deceptive reviews related to published information have exploded. These deceptive reviews are mostly conversation-based and laden with misinformation, which can spread quickly. The question-answering dialogue may be more vulnerable to being manipulated since platform users can answer the question with multiple facts or speculative and vague expressions(Sarrouti et al., 2021) that deliberately distribute misinformation. To improve the robustness of fact verification systems, they must also be valid for verifying the claims in question-answering dialogues.

There are two challenges to verifying the factual correctness of claims in question-answering dialogue. One is that claims in dialogue are often informal, sparse in factual content(Gupta et al., 2021), and contain personal opinions, slang, and colloquialisms, making it difficult to distinguish them from factual information. The other challenge is the difficulty of taking full advantage of the rich information contained in questions to support fact verification, such as specific domains, time frames, geographies, keywords, and descriptions. Regrettably, previous research has not sufficiently utilized the potential inherent in the questions posed by users. It is essential to recognize that the information in the questions plays a key role in facilitating deeper interactions among the various types of data during the verification process. To illustrate this point, as exemplified in the first case in Figure 1, question information can serve as a filtering mechanism to filter out noise and refine the information that is directly relevant to the verification task, e.g. *"the most common were social distancing and lockdown"* and *"the measures can slow down the outbreak"* in evidence, and *"social distancing"* and *"slowing the spread of the flu"* in claim. Besides, claim and evidence also have the same effect on the question, e.g. *"ibuprofen (advil) worsen COVID-19"* in the

*Corresponding author

LREC-COLING 2024, pages 10488–10498

second question. In addition, the majority of previous works for fact verification rarely considered label information. However, reasonable utilization of labels is beneficial to the fact verification process. Taking the consideration above, we investigated how to maximize the utilization of question for in-depth interaction among various types of information, while concurrently exploring the integration of label information within the verification process.

To meet these demands, this paper presents a novel framework, called **LI**[4] - **L**abel-**I**nfused **I**terative **I**nformation **I**nteracting, for question-answering dialogue based fact verification. The LI[4] consists of precisely engineered IIRF and FLEM components. Inspired by the Interactive Gating Mechanism (IGM) introduced in DABERTa(Sundriyal et al., 2022), we propose IIRF, which is designed to facilitate deep semantic interactions among question, evidence, and claim, aiming to emphasize key fragments within these three components while filtering out irrelevant noise. The FLEM easily and efficiently embed label information, encompassing categories such as "SUPPORTED," "REFUTED," and "NOTENOUGH-INFO," into the final feature representation. This incorporation is accomplished through the utilization of a specifically constructed prefix that assumes a distinctive prompting function, significantly mitigating the opaqueness associated with conventional black-box approaches to fact verification.

In summary, our contributions are as follows:

- We investigate an iterative interaction method to filter out the noise of question and evidence and refine information valuable for claim verification, achieving the purpose of more effective use of question and obtaining better verification performance.
- We innovatively implement label embedding technology for fact verification, aiming to strengthen the understanding ability of the model towards labels.
- We conduct the extensive experimental evaluation for the proposed model on three popular benchmark datasets, and the result shows LI[4] makes great progress on all experimental benchmarks and achieves new state-of-the-art performance[1].

2. Related Work

Fact Verification. In recent years, fact verification has gained significant attention in the NLP research community as a means of combating misinformation and disinformation. Previous studies on the fact verification are mainly focused on Wikipedia articles(Rashkin et al., 2017; Bekoulis et al., 2021; Wang et al., 2022a; Fajcik et al., 2022; Pan et al., 2023) and table-based verification(Chen et al., 2019; Shi et al., 2021; Liu et al., 2021; Zhou et al., 2022; Gu et al., 2022). However, fact verification in question-answering dialogue is still a preliminary exploratory stage. Several works explored fact verification for the dialogue context, mainly focusing on constructing datasets(Gupta et al., 2021; Zheng et al., 2022; Li et al., 2023). Meanwhile, several works have used question-answering dialogue data to construct fact verification benchmarks(Chen et al., 2021; Sarrouti et al., 2021; Park et al., 2021; Kim et al., 2021). There are few methods or frameworks specifically designed to solve fact verification in question-answering dialogue. Zou et al. (2023) propose DECKER, a commonsense QA fact verification model that is capable of bridging heterogeneous knowledge by uncovering latent relationships between structured and unstructured knowledge. But this is not suitable for the field we are exploring, because the question in commonsense QA is the content we want to judge whether or not, similar to a claim, rather than auxiliary information. Wang et al. (2022b) propose QaDialMoE, to our best knowledge, this is currently the only approach to investigate a question-answering di-

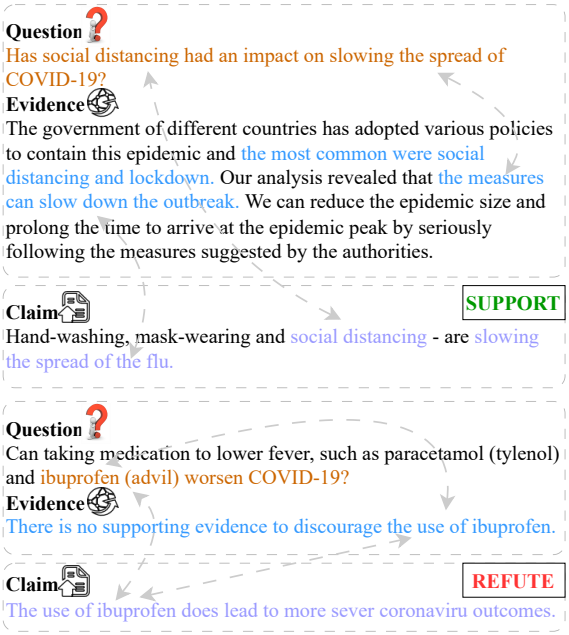

Figure 1: Question-answering dialogue based fact verification examples in the HEALTHVER dataset. Each example consists of a question, evidence, and a claim, where the claim is verified by evidence and question.

[1]Code and data are available at https://github.com/zxc123cc/LI4

alogue based on fact verification. To extend this effort, we propose a neural network, that focuses on various interactions among claims, questions, and evidence, while embedding knowledge of labels.

Information refining and filtering are very important in fact verification. In particular, evidence usually contains a lot of noise. This phenomenon is more serious after adding questions as additional input text. How to remove this noise and refine information that is strongly related to the claim is worth exploring. Previous research on information refining and filtering for fact verification tasks mainly focuses on evidence distillation(Nie et al., 2019; Ma et al., 2019; Wu et al., 2021; Yang et al., 2022) or claim span identification(Sundriyal et al., 2022). However, they only focused on the claim or evidence without considering it jointly. Sundriyal et al. (2022) uses Interactive Gating Mechanism(IGM) to further distinguish salient tokens inclusive in claim spans. Inspired by IGM, we improve it to IIRF for question-answering dialogue based fact verification. This method focuses on the interaction among question, evidence, and claim, and can control the degree of refinement and filtering by explicitly adjusting the number of iterations.

Label knowledge embedding. Label knowledge embedding can improve classification performance while maintaining almost the same computational cost(Zhang et al., 2017; Du et al., 2019). Xiong et al. (2021) splices the label with the original input and then takes the [CLS] representation of BERT(Devlin et al., 2018) for classification, fully tapping the potential of BERT. Liu et al. (2022) propose a collaborative attention network with label embeddings(CNLE) that co-encodes text and labels into their mutually participating representations. In this paper, we leverage label embedding technology to integrate the rich information contained in labels into the feature representation so as to improve the performance of fact verification.

3. Task Statement

In the question-answering dialogue based fact verification, a sample $\mathcal{S}$ is defined as $\mathcal{S}$ = (c, (e, q), y), representing a given claim c, evidence e, the corresponding question q, and the associated label y. The goal is to verify the factuality of claim c by the given evidence-question pair (e, q). Each claim c is associated with a veracity y taking one of the class labels from {SUPPORTED,REFUTED,$\cdots$}. Beyond the label as SUPPORTED or REFUTED, the classification task has one more label called NEUTRAL or NOT ENOUGH INFO (NEI), which means the current evidence is insufficient to verify the claim. Then, the task changed from a 2-way to a 3-way classification task. Unlike most fact verification processes, which predominantly rely on the evidence alone, this task carries an additional question. It is worth noting that the question is able to provide a great deal of additional information. Leveraging the question reasonably can improve prediction performance, which is key to this task.

4. Approach

In this section, we introduce our novel LI[4] model, which considers the interaction among question, evidence, and claim, while concurrently embedding label knowledge. Figure 2 shows the overall model architecture. Our method consists of three components: the feature extractor (§4.1) with a transformer encoder backbone, the iterative information refining and filtering module (§4.2) for interaction among different parts of the input, and the fact label embedding module (§4.3) for infusing label knowledge.

4.1. Feature Extractor

The feature extractor takes claim-question-evidence pair as input text, which are subsequently fed into a transformer encoder to learn their joint semantics representations.

4.1.1. Input Process

For each sample, we splice the claim-question-evidence pair. Then we prefixed the sentences and split each part with $\langle/s\rangle$. This prefix serves the purpose of incorporating label information. To feed into the encoder, we need to perform tokenization of each sentence. Specifically, we tokenize the prefix and claim-question-evidence pair p, c, q, e into four token sequences denoted as **P**, **C**, **Q** and **E**. The sequence input $S_{p,c,e,q} = [\langle s\rangle, \mathbf{P}, \langle/s\rangle, \mathbf{C}, \langle/s\rangle, \mathbf{E}, \langle/s\rangle, \mathbf{Q}, \langle/s\rangle]$. where $\langle s\rangle$ and $\langle/s\rangle$ are delimiters indicating the beginning and end of each token sequence. In order to ensure the consistency and structural integrity of each part of the information, when the length of a given text surpasses the prescribed maximum threshold, the evidence and questions are subject to an equitable truncation, thereby preventing an excessive loss of information within a singular segment.

4.1.2. Joint Semantics Representation

We feed the joint token sequence into a transformer-based encoder to learn the contextualized semantics representation:

$$H = Encoder(S_{p,c,e,q}) \quad (1)$$

where $H \in \mathbb{R}^{n\times d}$ denotes the learned joint semantics representation. Here n is the maximum length

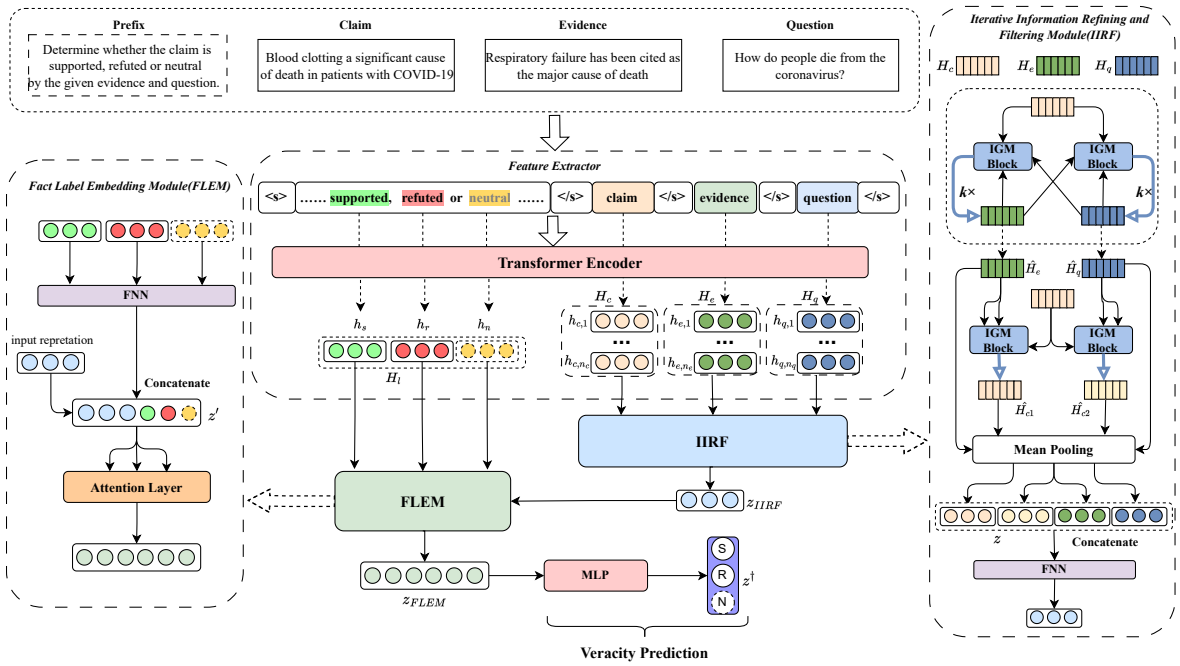

Figure 2: Model Architecture Overview of LI[4]. The prefix, claim, evidence, and question are concatenated and then fed into a transformer-based encoder to obtain the joint semantics representations. The Iterative Information Refining and Filtering Module (IIRF) is designed to iteratively filter out the noise of question and evidence, and refine claim information and the Fact Label Embedding Module(FLEM) is used to embed label information.

of input and d is the representation vector dimension.

Subsequently, we extract $H_c \in \mathbb{R}^{n_c \times d}$,$H_q \in \mathbb{R}^{n_q \times d}$,$H_e \in \mathbb{R}^{n_e \times d}$, the representations corresponding to claim, question and evidence from H. Here n_c, n_q and n_e are their corresponding lengths respectively. Simultaneously, we extract label features $H_l \in \mathbb{R}^{n_l \times d}$, here n_l denotes the number of labels. In cases where the labels contain multiple words, we take the mean-pooling of the token features to obtain the label-aware feature vector.

4.2. Iterative Information Refining and Filtering Module (*IIRF*)

In order to filter out the noise of question and evidence, and refine the information of claim, inspired by IGM(Sundriyal et al., 2022), we propose *Iterative Information Refining and Filtering Module*. Unlike previous IGM, the inputs accepted by the conflict gate and refine gate of the IGM module we use here may be different. To begin with, the inputs $H_1 \in \mathbb{R}^{n_1 \times d}$, , $H_2 \in \mathbb{R}^{n_2 \times d}$ and $H_3 \in \mathbb{R}^{n_3 \times d}$ are max pooled to obtain $H_{p_1}, H_{p_2}, H_{p_3} \in \mathbb{R}^d$. Here n_1, n_2 and n_3 represent their corresponding lengths. These vectors are passed through a series of gates. The first is *conflict gate* C, which is used to capture semantic features in H_1 that conflict with H_2.

$$\mu_c = \sigma(H_{p_1} W_{c_1} + H_{p_2} W_{c_2} + b_{c_1}) \quad (2)$$

$$C = tanh(H_{p_1} \odot \mu_c W_{c_3} + H_{p_2} \odot (1 - \mu_c) W_{c_4} + b_{c_2}) \quad (3)$$

The *refine gate* R, on the other hand, is aimed at capturing the semantically similar features between H_1 and H_3.

$$\mu_r = \sigma(H_{p_1} W_{r_1} + H_{p_3} W_{r_2} + b_{r_1}) \quad (4)$$

$$R = tanh(H_{p_1} \odot \mu_r W_{r_3} + H_{p_3} \odot \mu_r W_{r_4} + b_{r_2}) \quad (5)$$

The last of them being the *adaptive gate* A, is used to retain maximum differential information from C and R, thereby filtering the noise that is not related to H_2 and refining the semantic representations similar to H_3.

$$A = R + (1 - \mu_r) \odot C \quad (6)$$

$$\hat{H} = tanh(A W_a + b_a) \odot H_1 \quad (7)$$

After passing these three gates, we obtain the vector $\hat{H}$ that has been refined and filtered.

In the question-answering dialogue based fact verification task, the claim is the target we want to classify based on question and evidence. That is, the question and evidence should refine the information from the claim while filtering information that conflicts with each other.

We first perform information refining and filtering on e and q. It is noteworthy that this process is iterable, that is, the outputs H_q and H_e of the previous round can serve as the input of the current round. Let H^i be the output of the i^{th} round of iteration. The iteration process is given by Equation 8 and 9.

$$H_e^i = IGM(H_e^{i-1}, H_q^{i-1}, H_c) \tag{8}$$

$$H_q^i = IGM(H_q^{i-1}, H_e^{i-1}, H_c) \tag{9}$$

After k rounds of iterations, we can get the final representations of the question and evidence, where k is a hyperparameter representing the maximum number of iterations:

$$\hat{H}_e = H_e^k, \hat{H}_q = H_q^k \tag{10}$$

Next, we use $\hat{H}_e$ and $\hat{H}_q$ to refine and filter the characteristics of the claim respectively, which will make the claim focus on different information of question and evidence.

$$\hat{H_{c_1}} = IGM(H_c, \hat{H}_e, \hat{H}_e) \tag{11}$$

$$\hat{H_{c_2}} = IGM(H_c, \hat{H}_q, \hat{H}_q) \tag{12}$$

At the end of the module, $\hat{H_{c_1}}$, $\hat{H_{c_2}}$, $\hat{H}_e$ and $\hat{H}_q$ are mean pooled to obtain $z_{c_1}, z_{c_2}, z_e, z_q \in \mathbb{R}^d$. Subsequently, they are concatenated and passed through a Feed-Forward Network(FNN) to obtain the final representation.

$$z = Concat(z_{c_1}, z_{c_2}, z_e, z_q) \tag{13}$$

$$z_{IIRF} = FNN(z) \tag{14}$$

4.3. Fact Label Embedding Module (*FLEM*)

The Fact Label Embedding Module is proposed to enhance the performance of fact verification by integrating label-related information and strengthening the understanding ability of the classification model towards labels. First, we construct a prefix to introduce label information. As shown in Figure 2, the prefix *"Determine whether the claim is supported, refuted or neutral by the given evidence and question"* introduces three labels: "supported", "refuted" and "neutral" (some datasets do not contain neutral label). Their features are represented as h_s, h_r, h_n respectively, and the combination is H_{n_l}. Subsequently, we map H_l using a FNN layer.

$$H_l^{'} = FNN(H_l) \tag{15}$$

where $H_l^{'} = [h_{l,1}^{'}; h_{l,2}^{'}; ...; h_{l,n_l}^{'}]$, $h_{l,i}^{'} \in \mathbb{R}^{d^{'}}$ denotes representation of the i^{th} label after dimensionality reduction. Here $d^{'}$ is the dimension after dimensionality reduction.

Next, to integrate the label features with the input representation, we concatenate them into a new vector.

$$z^{'} = Concat(z_{IIRF}, h_{l,1}^{'}, h_{l,2}^{'}, ..., h_{l,n_l}^{'}) \tag{16}$$

where $z^{'} \in \mathbb{R}^{l \times d^{'} + d}$ is the concatenated vector. To better embed the corresponding label features, we apply an attention mechanism to select the concatenated vector. Specifically, we use the attention mechanism to calculate a weight vector that represents the importance of the input representation and label features. Then the weighted sum of the input representation and label features based on the weight vector to obtain the final embedded representation z_{FLEM}.

$$[Q, K, V] = z^{'} W_{QKV} \tag{17}$$

$$z_{FLEM} = softmax(\frac{QK^T}{\sqrt{l \times d^{'} + d}})V \tag{18}$$

where $W_{QKV} \in \mathbb{R}^{3 \times d}$ is the projection matrix. By employing this approach, the Fact Label Embedding Module effectively integrates label-related information with feature information, thereby enhancing the understanding and classification capability of the classification model towards labels and improving the performance of fact verification.

4.4. Veracity Prediction and Model Training

Finally, z_{FLEM} is fed into a multi-layer perceptron (MLP) layer, and the softmax function is applied to predict the veracity label as follows:

$$z^{\dagger} = MLP(z_{FLEM}) \tag{19}$$

$$\hat{y}_i = softmax(z_i) = \frac{e^{z_i}}{\sum_{z_i \in z^{\dagger}} e^{z_i}} \tag{20}$$

For model training, we minimize the loss $\mathcal{L}_{ce}$ as follows:

$$\mathcal{L}_{ce} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log \hat{y}_i \tag{21}$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss for fact verification task. N denotes the number of samples. y_i and $\hat{y}_i$ denote the gold and predicted label of the i^{th} claim, respectively.

5. Experimentation

5.1. Dataset and Evaluation Metrics

To evaluate our method, we carry out detailed experiments on three English benchmark datasets:

HEALTHVER(Sarrouti et al., 2021), FAVIQ(Park et al., 2022), and COLLOQUIAL(Kim et al., 2021).

HEALTHVER HEALTHVER is an evidence-based fact verification dataset, it's primarily used for answering medical-related questions and includes various questions and corresponding answers in the field of medicine. The relationship between each evidence statement and its corresponding claim is manually annotated as SUPPORT, REFUTE, or NEUTRAL. The dataset consists of 14,330 evidence-claim pairs. We evaluate the performance of our model on the HEALTHVER dataset using metrics macro precision, macro recall, macro F1-score, and accuracy.

FAVIQ FAVIQ is constructed from information-seeking questions (Kwiatkowski et al., 2019) and their ambiguities (Min et al., 2020), which focuses on answering questions and queries about common topics, including social media, travel, sports, and more. It consists of the A set and the R set. The A set is the main dataset, consisting of 25,956 claims that have been disambiguated by transforming ambiguous questions into claim form. The R set is an additional dataset, containing 162,420 claims extracted and generated from regular question-answer pairs. Most instances include the question, and evidence from Wikipedia paragraphs, and are annotated as SUPPORT or REFUTE. For the A set, only the A development set is used, and no questions are generated for the A test set. The evaluation metric used for model performance is accuracy.

COLLOQUIAL In order to investigate how fact verification systems behave on colloquial claims, COLLOQUIAL is constructed by transferring the styles of claims from FEVER into colloquialism. Existing fact verification systems that perform well on claims in formal style significantly degenerate on colloquial claims with the same semantics. Our model can effectively deal with the challenge of colloquialism due to the filter function in IIRF. This dataset does not contain questions, so we use the question synthesized by (Wang et al., 2022b). Finally, each instance in COLLOQUIAL consists of a synthetic question, an evidence from Wikipedia passages, and a claim with the label SUPPORTED, REFUTED, or NEI. We use the label accuracy as our evaluation metric.

5.2. Experimental Settings

Our implementations are based on the public Pytorch implementation from Transformers[2]. We leverage RoBERTa-Large (Liu et al., 2019) to initialize the parameters of the encoder. The maximum input length and dimension of label information are set to 512 and 128 respectively. The maximum number of iterations k is set to 1 on the COLLOQUIAL dataset and to 3 for the remaining datasets. We apply AdamW optimizer (Loshchilov and Hutter, 2017) in training with a learning rate 2e-5. We set the batch size to 32 on the HEALTHVER dataset. Gradient accumulation is turned on for other datasets and the batch size is set to 12. The gradient accumulation steps for FAVIQ R and the remaining datasets are 4 and 3 respectively. All models are fine-tuned on a single Tesla A100 GPU with 40 GB memory.

[2]https://github.com/huggingface/transformers

5.3. Overall Verification Results

We compare our proposed model LI^4 with other methods on the HEALTHVER, FAVIQ, and COLLOQUIAL. The experimental results are shown in Table 1, Table 3, and Table 4 respectively.

HEALTHVER Based on the question, evidence, and claim in the original dataset, we evaluate the performance of our method on HEALTHVER. As shown in Table 1, LI^4 outperforms all previous methods by a large margin. Our method has an accuracy of 85.41% in the test set of HEALTHVER, achieving a new state-of-the-art on the dataset, which is 1.15% higher than the previous best method(QaDialMoE) and outperforms it with improvements of 0.95% in macro precision, 1.44% in macro recall, and 1.25% in macro F1-score.

Models	P	R	F1	Acc
BERT-base	73.45	73.70	73.54	74.82
SciBERT	76.62	78.15	77.12	78.11
BioBERT	74.07	75.73	74.59	76.52
T5-base	80.82	79.00	79.60	80.69
BART-large	81.48	81.20	81.33	82.23
RoBERTa-large	82.24	81.48	81.78	82.72
QaDialMoE	83.95	82.83	83.29	84.26
Ours	**84.90**	**84.27**	**84.54**	**85.41**

Table 1: Comparative performance on HEALTHVER test set.

FAVIQ The methods of obtaining evidence on FAVIQ can be divided into three types: DPR(Qu et al., 2020), evidentiality-guided generator(EG)(Asai et al., 2021), and positive evidence(PE). We only evaluate our approach on PE. The positive evidence is the top passage that contains the answer to the original question which is retrieved by TF-IDF(Park et al., 2021).

As shown in Table 2, PE contains a certain proportion of N/A fields, which means that the evidence is empty and will have a great impact on classification performance. Table 3 presents the comparative performance of FAVIQ. For A set, our methods achieve a new state-of-the-art with an accuracy of 79.0%. For the R set, our methods with positive evidence can reach remarkable performances

		Total	N/A	ratio(%)
Train	A	17,008	6,084	35.77
	R	140,977	43,536	30.88
Dev	A	4,260	1,472	34.55
	R	15,566	4,853	31.18
Test	A	4,688	-	-
	R	5,877	1,884	32.06

Table 2: Statistics of N/A ratio in FAVIQ dataset, consisting of A set and R set.

with 86.0% on the dev set and 86.5% on the test set. The experimental results on the dev set are almost the same as the previous best methods, while on the test set our method outperforms prior methods, achieving significant improvements with 0.5% (86.5% vs. 86.0%). Compared with other datasets, the improvement of LI[4] on the FAVIQ is not obvious. We analyze that this is because the label of claim is strongly related to evidence, and N/A prevents a large number of claims from being effectively verified, and the iterative interaction part in IIRF does not play its full role.

Models	A-dev	R-dev	R-test
Claim only BART	51.0	59.4	59.4
TF-IDF + BART	65.1	74.2	71.2
DPR + BART	66.9	76.8	74.6
FiD(base)	67.8	-	-
FiD + EG	69.6	-	-
QaDialMoE + DPR	70.8	78.0	75.3
QaDialMoE + EG	74.9	-	-
QaDialMoE + PE	78.7	**86.1**	86.0
Ours + PE	**79.0**	86.0	**86.5**

Table 3: Fact verification accuracy on FAVIQ.

COLLOQUIAL Since COLLOQUIAL does not have original questions, we use the synthetic questions that Wang et al. (2022b) have generated using a prompt module. As shown in Table 4, our approach obtains a new state-of-the-art label accuracy and achieves a remarkable improvement of 1.9% (91.4% vs. 89.5%). Colloquial claims tend to include filter words (e.g.,*"yeah"*, *"you know"*), comments, or personal opinions that do not require verification. In addition, questions generated using the prompt module also contain a lot of noise compared to manually annotated questions. Fortunately, our IIRF can effectively deal with the above challenges by filtering noisy information and therefore achieves great improvements compared to other datasets.

5.4. Ablation Study

To further investigate each component's contribution to the whole network, we perform ablation studies by deliberately removing certain modules and comparing the results on HEALTHVER dataset. The overall configuration remains consistent, while only the module under investigation is removed from the whole network. The results are shown in Table 5.

Models	Document Retrieval +Evidence Selection	Acc
KGAT(BERT)	DPR + BERT	51.2
	WikiAPI + BERT	53.2
	Evidence Oracle	57.3
KGAT (CorefBERT)	DPR + BERT	61.0
	WikiAPI + BERT	60.9
	Evidence Oracle	67.7
QaDialMoE	Evidence Oracle	89.5
Ours	Evidence Oracle	**91.4**

Table 4: Fact verification label accuracy on COLLOQUIAL.

w/o IIRF: We conduct an ablation study on the HEALTHVER dataset without IIRF. As presented in Table 5, LI[4] has dropped by 1.81% (85.41% vs. 83.60%) and 1.77%(84.54% vs. 82.77%) in accuracy and macro F1-score. The drop in the performance across all metrics by a large margin, fully illustrates the effectiveness of IIRF. Meanwhile, the impact of the iterations in IIRF will be discussed in more detail in Sec. 5.5.

w/o FLEM: When FLEM is removed, we observe a drop in performance across all the metrics. Our model has dropped by 0.33% (85.41% vs. 85.08%) and 0.27%(84.54% vs. 84.27%) in accuracy and macro F1-score. This shows that introducing label information can enhance the representation ability of each category and further improve the performance of fact verification tasks. More specific details will be discussed in Sec. 5.6. In addition, it is worth noting that when FLEM is removed, that is, only the IIRF component is retained, our method still outperforms all previous methods and achieves a new state-of-the-art result, which further illustrates the effectiveness of IIRF.

Models	P	R	F1	Acc
Complete model	**84.90**	**84.27**	**84.54**	**85.41**
- w/o IIRF	83.21	82.48	82.77	83.60
- w/o FLEM	85.07	83.73	84.27	85.08

Table 5: Ablation study on HEALTHVER test set.

5.5. Effect of Iterations in IIRF

To further understand the effectiveness of the proposed IIRF, we investigate the impact of the number of iterations. Figure 3 reports accuracy on four datasets when using different values for the number

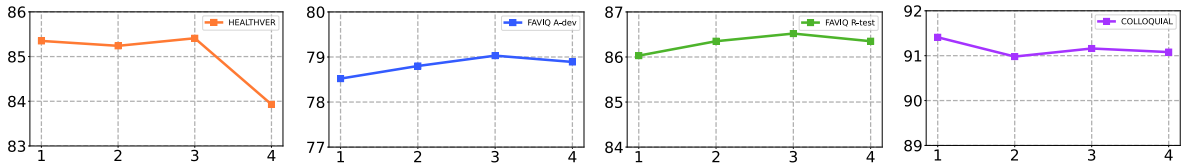

Figure 3: Sensitivity analysis of the number of Conflict/Refine iterations between question and evidence. The horizontal and vertical axes represent the number of iterations and accuracy respectively.

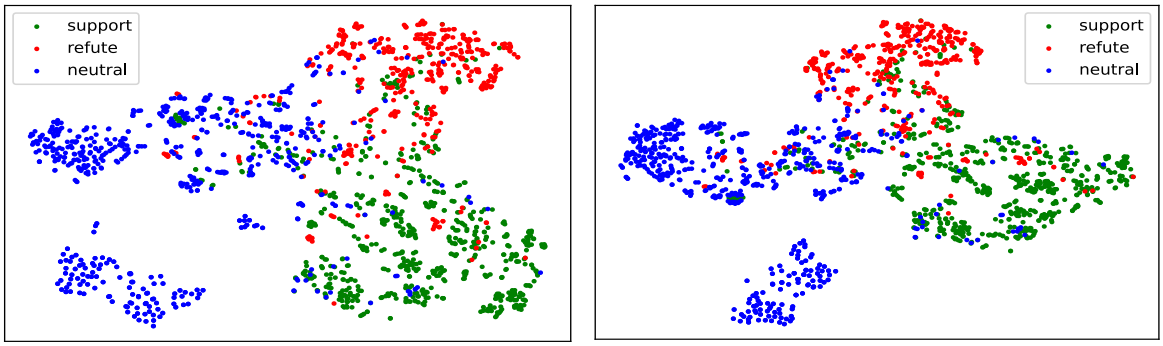

Figure 4: t-SNE projection before and after adding FLEM on the HEALTHVER dataset.

of iterations. We can observe that as the number of iterations increases, except for the COLLOQUIAL dataset, the performance of other datasets shows a trend of rising first and then falling. On the COLLOQUIAL dataset, the number of iterations that achieved the best performance was 1, while on other datasets it was 3. This is consistent with our initial conjecture. In the COLLOQUIAL dataset, the overall length of question, evidence, and claim is shorter, the annotation quality is higher, and there is not much noisy information, so only one iteration is needed, while other datasets contain a lot of long evidence or even missing evidence, only a small part of the information is valuable to judge the classification results. Experiments showed a drop in the fourth iteration on all datasets. We analyzed that too many iterations will cause over-fitting, thus reducing performance.

5.6. Quality of Representation

We adopt t-SNE(Van der Maaten and Hinton, 2008) to project representations of the hidden state of each sample before final classification into a 2-dimensional space. From the result shown in Figure 4, we can see that representations of using FLEM (Figure 4(b)) and not using FLEM (Figure 4(a)) have a clear distinction, showing that FLEM is able to produce high-quality representations for Fact Verification. Specifically, we can observe that the representation of each category using FLEM is more compact, and the division among categories is more obvious. Interestingly, the representation of NEUTRAL is divided into two parts, which indicates that there may be some connection between the two parts of the sample.

Here, we conduct a case study on focused on the neutral data subset of HEALTHVER. We find that the data in the farthest cluster is very easy to classify because its claim and evidence are almost irrelevant, such as the claim *"There is much more to coronaviruses than SARS-CoV-2. Coronaviruses are actually a family of hundreds of viruses"* and evidence *"Recent research results suggest that bats or pangolins might be the original hosts for the virus based on comparative studies using its genomic sequences"*. In contrast, the data of clusters close to SUPPORT and REFUTE are hard samples. Such as the claim *"Common Steroid Could Be Cheap and Effective Treatment for Severe COVID-19"* and evidence *"Many drugs have shown promise for the treatment of COVID-19"*. In this case, we cannot infer whether *"Many drugs"* contain a *"Common Steroid"*, yet *"Could Be"* makes the sentence ambiguous. Strengthening the model's ability to judge these samples is an important way to improve performance.

6. Conclusion

In this study, we introduce the LI[4] neural network designed for the task of fact verification in question-

answering dialogue, demonstrating superior verification performance across three popular datasets: HEALTHVER, FAVIQ, and COLLOQUIAL. The core focus of LI^4 centers on fostering comprehensive interactions among questions, evidence, and claims, coupled with the incorporation of label information. LI^4 exhibits a noteworthy capacity for robust generalization across diverse datasets, benefiting from its iterative mechanism, which readily adapts to both complex and straightforward questions and evidence. Ablation studies and in-depth analyses provide further evidence of the efficacy of both the constituent components and the iterative mechanisms within the LI^4 model. We hope our work can facilitate fact verification in the question-answering dialogue domain, strengthen the full use of the question so that it further interacts with evidence and claims, and consider label information to make the verification process more explicit.

7. Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant Nos.62272132).

8. Bibliographical References

Akari Asai, Matt Gardner, and Hannaneh Hajishirzi. 2021. Evidentiality-guided generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2112.08688*.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. A review on fact extraction and verification. *ACM Comput. Surv.*, 55(1).

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can nli models verify qa systems' predictions? *arXiv preprint arXiv:2104.08731*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Cunxiao Du, Zhaozheng Chen, Fuli Feng, Lei Zhu, Tian Gan, and Liqiang Nie. 2019. Explicit interaction model towards text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6359–6366.

Martin Fajcik, Petr Motlicek, and Pavel Smrz. 2022. Claim-dissector: An interpretable fact-checking system with joint re-ranking and veracity prediction. *arXiv preprint arXiv:2207.14116*.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders nlp fact-checking unrealistic for misinformation. *arXiv preprint arXiv:2210.13865*.

Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. 2022. Pasta: table-operations aware fact verification via sentence-table cloze pre-training. *arXiv preprint arXiv:2211.02816*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Dialfact: A benchmark for fact-checking in dialogue. *arXiv preprint arXiv:2110.08222*.

Byeongchang Kim, Hyunwoo Kim, Seokhee Hong, and Gunhee Kim. 2021. How robust are fact checking systems on colloquial claims? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1535–1548.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Helma: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Minqian Liu, Lizhao Liu, Junyi Cao, and Qing Du. 2022. Co-attention network with label embedding for text classification. *Neurocomputing*, 471:61–69.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6859–6866.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Faviq: Fact verification from information-seeking questions. *arXiv preprint arXiv:2107.02153*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512.

Qi Shi, Yu Zhang, Qingyu Yin, and Ting Liu. 2021. Logic-level evidence retrieval and graph-based verification network for table-based fact verification. *arXiv preprint arXiv:2109.06480*.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Empowering the fact-checkers! automatic identification of claim spans on twitter. *arXiv preprint arXiv:2210.04710*.

Sander van Der Linden, Jon Roozenbeek, and Josh Compton. 2020. Inoculating against fake news about covid-19. *Frontiers in psychology*, page 2928.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Hao Wang, Yangguang Li, Zhen Huang, and Yong Dou. 2022a. Imci: Integrate multi-view contextual information for fact extraction and verification. *arXiv preprint arXiv:2208.14001*.

Longzheng Wang, Peng Zhang, Xiaoyu Lu, Lei Zhang, Chaoyang Yan, and Chuang Zhang. 2022b. Qadialmoe: Question-answering dialogue based fact verification with mixture of experts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3146–3159.

Lianwei Wu, Yuan Rao, Ling Sun, and Wangbo He. 2021. Evidence inference networks for interpretable claim verification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14058–14066.

Yijin Xiong, Yukun Feng, Hao Wu, Hidetaka Kamigaito, and Manabu Okumura. 2021. Fusing label embedding into bert: An efficient improvement for text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1743–1750.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. *arXiv preprint arXiv:2209.14642*.

Honglun Zhang, Liqiang Xiao, Wenqing Chen, Yongkun Wang, and Yaohui Jin. 2017. Multi-task label embedding for text classification. *arXiv preprint arXiv:1710.07210*.

Chujie Zheng, Jinfeng Zhou, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zhengyu Niu, Hua Wu, and Minlie Huang. 2022. Cdconv: A benchmark for contradiction detection in chinese conversations. *arXiv preprint arXiv:2210.08511*.

Yuxuan Zhou, Xien Liu, Kaiyin Zhou, and Ji Wu. 2022. Table-based fact verification with self-adaptive mixture of experts. *arXiv preprint arXiv:2204.08753*.

Anni Zou, Zhuosheng Zhang, and Hai Zhao. 2023. Decker: Double check with heterogeneous knowledge for commonsense fact verification. *arXiv preprint arXiv:2305.05921*.

9. Language Resource References

Kim, Byeongchang and Kim, Hyunwoo and Hong, Seokhee and Kim, Gunhee. 2021. *How Robust are Fact Checking Systems on Colloquial Claims?* PID https://github.com/bckim92/colloquial-claims.

Park, Jungsoo and Min, Sewon and Kang, Jaewoo and Zettlemoyer, Luke and Hajishirzi, Hannaneh. 2022. *FaVIQ: Fact Verification from Information seeking Questions*. PID https://faviq.github.io.

Sarrouti, Mourad and Abacha, Asma Ben and M'rabet, Yassine and Demner-Fushman, Dina. 2021. *Evidence-based Fact-Checking of Health-related Claims*. PID https://github.com/sarrouti/healthver.