QaDialMoE: Question-answering Dialogue based Fact Verification with Mixture of Experts

Longzheng Wang[1,2], **Peng Zhang**[2,3]*, **Xiaoyu Sean Lu**[3], **Lei Zhang**[1,2], **Chaoyang Yan**[1,2], **Chuang Zhang**[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]School of Cyber Security, Nanjing University of Science and Technology
{wanglongzheng,pengzhang,zhanglei0510,yanchaoyang,zhangchuang}@iie.ac.cn
xiaoyu.lu@njust.edu.cn

Abstract

Fact verification is an essential tool to mitigate the spread of false information online, which has gained a widespread attention recently. However, a fact verification in the question-answering dialogue is still underexplored. In this paper, we propose a neural network based approach called question-answering dialogue based fact verification with mixture of experts (QaDialMoE). It exploits questions and evidence effectively in the verification process and can significantly improve the performance of fact verification. Specifically, we exploit the mixture of experts to focus on various interactions among responses, questions and evidence. A manager with an attention guidance module is implemented to guide the training of experts and assign a reasonable attention score to each expert. A prompt module is developed to generate synthetic questions that make our approach more generalizable. Finally, we evaluate the QaDialMoE and conduct a comparative study on three benchmark datasets. The experimental results demonstrate that our QaDialMoE outperforms previous approaches by a large margin and achieves new state-of-the-art results on all benchmarks. This includes the accuracy improvements on the HEALTHVER as **84.26%**, the FAVIQ A dev set as **78.7%**, the FAVIQ R dev set as **86.1%**, test set as **86.0%**, and the COLLOQUIAL as **89.5%**. To our best knowledge, this is the first work to investigate a question-answering dialogue based fact verification, and achieves new state-of-the-art results on various benchmark datasets.[1]

1 Introduction

Fact Verification, aiming to validate the factuality of claims against a corpus of documents, is an important NLP area (Cohen et al., 2011) and has been explored to various applications such as detecting fake news, rumor, and deceptive opinions (Rashkin et al., 2017; Thorne et al., 2018; Goodrich et al., 2019; Vaibhav et al., 2019; Chen et al., 2020; Kryscinski et al., 2020). The majority of existing research focuses on media such as news, tables and Wikipedia passages (Guo et al., 2021; Bekoulis et al., 2021), while rarely consider the fact verification in the question-answering dialogue. In a dialogue safety domain, related studies either focus on enabling dialogue agents to resist adversarial attacks (Dinan et al., 2019a) or on forestalling aggressive or biased responses from dialogue agents (Henderson et al., 2018; Sap et al., 2019; Xu et al., 2020).

Question: Can animals spread COVID-19 to people?
Evidence: There is evidence that SARS-CoV-2 can infect felines, dogs and minks, and there is evidence of human-to-animal infection.

Response with Label:
No proof yet that pets can get COVID-19 from owners. [REFUTED]

Question: Who wrote the song "this is me" and "rewrite the stars"?
Evidence: The Greatest Showman is a 2017 American musical biographical drama film directed by Michael Gracey in his directorial debut …original songs from Benj Pasek and Justin Paul…

Response with Label:
Benj Pasek and Justin Paul wrote the song "this is me" and "rewrite the stars" for the film the greatest showman. [SUPPORTED]

Figure 1: Two examples of question-answering dialogue based fact verification. The first response is retrieved from a real-world question about COVID-19. The second response is derived from a QA corpus using an ambiguous information-seeking question.

However, misinformation online can spread quickly and cause public health crises due to the abuse of dialogue agents, especially questions about the pandemic of COVID-19 (Naeem et al., 2021). The first example in Figure 1 shows a popular question about COVID-19 asked by information

*Corresponding author.
[1]Code and data are available at https://github.com/wishever/QaDialMoE

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3146 - 3159
December 7-11, 2022

seekers online. The question-answering dialogue may be more vulnerable to be manipulated, since Internet users can answer the question with multiple facts or speculative and vague expressions (Sarrouti et al., 2021) that deliberately distribute misinformation. For improving the robustness of fact verification systems, they must also be valid for verifying the responses in question-answering dialogues.

The majority of previous works for the fact verification mainly focused on reasoning against pieces of evidence from Wikipedia passages, while rarely considered questions sought by Internet users. However, the questions also contain rich information to support the fact verification. Figure 1 shows two examples for the question-answering dialogue based fact verification, where the questions were posed by real users. We can see that the questions contain some critical parts (e.g., *"animals"*, *"people"*, *"who"*), which indicate the confusions of information seekers and the vulnerable part of responses. Taking the consideration above, we explore the fact verification in the question-answering dialogue and investigate how to exploit questions in the verification process.

In this paper, we present **QaDialMoE**, a neural network approach for **Q**uestion-**a**nswering **Dial**ogue based Fact Verification with **M**ixture **o**f **E**xperts. Inspired by that mixture of experts is applied in both dialogue systems (Le et al., 2016a) and fact verification fields (Zhou et al., 2022), we implement each expert with the same neural architecture to focus on different parts of inputs (e.g., the relationship between the response and the question). Specifically, to make our approach more generalizable, we propose a prompt module to generate *questions* in case that the original data only has *responses*. Then each expert takes the same feature as the input from the output of feature extractor module and learns to deal with the meaning of questions, responses and evidence. We design a management module to guide the training of experts by assigning a unique attention score to each expert, and combine their verification results efficiently. However, previous models tend to incorrectly predict a response as SUPPORTED when there is a significant overlap between the response and the evidence. Similarly, it incorrectly predict SUPPORTED or REFUTED for a NEI response because of the word overlap. Note that NEI is short for NOTENOUGHINFO. To alleviate this problem, we introduce an attention guidance module to generate a prior assumption and guide the manager paying more attention to the input part with few word overlap.

We conduct experiments on three benchmark datasets HEALTHVER (Sarrouti et al., 2021), FAVIQ (Park et al., 2022) and COLLOQUIAL (Kim et al., 2021). Experimental results demonstrate that our model outperforms previous systems by a large margin and achieves new state-of-the-art results on all of them. The main contributions of this paper can be summarized as follows:

- We explore a fact verification in the question-answering dialogue. To our best knowledge, this is the first study to investigate a question-answering dialogue based fact verification and to improve the applicability of fact verification systems.
- We introduce a use of mixture of experts and a manager with an attention guidance module for question-answering dialogue based fact verification, aiming to exploit questions and evidence efficiently in the verification process.
- We propose a prompt module to make our approach more generalizable, which can generate *questions* by given *responses*.
- Our approach achieves new state-of-the-art results on all experimental benchmarks, outperforming previous approaches by a large margin.

2 Task Background

In this paper, we study the task of question-answering dialogue based fact verification. Given a question Q, a response R and evidence E from Wikipedia passages, the goal is to verify the factuality of the response by the question and evidence with the label SUPPORTED or REFUTED. Beyond the label as SUPPORTED or REFUTED, the classification task has one more label called NEUTRAL or NEI, which means no enough information and cannot make a decision. Then, the task becomes a 3-way classification task.

Prior works have used question-answering dialogue data to create fact verification benchmarks (Demszky et al., 2018; Jiang et al., 2020; Pan et al., 2021; Chen et al., 2021; Sarrouti et al., 2021; Park et al., 2022). Most fact verification

processes only use evidence while rarely considering questions. However, we believe that the additional question that contains rich information is helpful to support the final prediction. In this study, we employ questions and evidence to validate the responses, which we formulate as the question-answering dialogue based fact verification task.

3 Methodology

In this section, we present our proposed framework QaDialMoE, that leverages a set of experts to simultaneously consider the meaning of questions, responses and evidence from Wikipedia passages. The overall model structure is illustrated in Figure 2. Our method consists of three components: the feature extractor (§3.1) with a prompt module and a transformer encoder backbone, the mixture of experts module (§3.2) for dealing with different parts of input, and the management module (§3.3) for guiding the training of experts and combining their ability of verification effectively.

3.1 Feature Extractor

In this section, a prompt module is proposed to generate *questions* by given *responses*. Subsequently, a transformer-based encoder parses the response-question (original or synthetic) -evidence pair and learns their joint semantics representations.

3.1.1 Prompt Module

Since question-answering dialogue based fact verification is still underexplored, few benchmark datasets use question-answering dialogue to retrieve or create responses (Sarrouti et al., 2021; Park et al., 2022). To make our approach more generalizable and explore the effectiveness of question-answering dialogue in the fact verification task, we propose a prompt module to generate *questions*. Specifically, we only use the *responses* in original data as the input. This can be easily generalized to more datasets. Then we leverage a question-generation model to synthesize *questions* by the given input. The synthetic questions are further passed to transformer encoder layers to learn response-question-evidence joint semantics (§3.1.2), and to an attention guidance module for generating prior assumptions (§3.3.1). In this paper, we implement it with T5 (Raffel et al., 2019), a transformer-based pre-trained model.

3.1.2 Joint Representation Learning

Given the response-question-evidence pair (§3.1.1), we construct a transformer-based encoder (Vaswani et al., 2017) to capture the joint semantics representation. Specifically, we tokenize the response-question-evidence pair r, q, e into three token sequences **R**, **Q** and **E**. Then the joint token sequence $\mathbf{L}_{r,q,e} = [\langle s\rangle, \mathbf{R}, \langle /s\rangle, \mathbf{Q}, \mathbf{E}, \langle /s\rangle]$, where $\langle s\rangle$ and $\langle /s\rangle$ are the separators that indicate the beginning and the end of each token sequence. Then we feed the joint token sequence into a transformer-based encoder to learn the contextualised representation embedding:

$$\mathbf{H} = f_{LM}(\mathbf{L}_{r,q,e}) \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{n\times d}$ denotes the learned joint semantics representation. Here n is the maximum length of input and d is the representation vector dimension. f_{LM} refers to the joint representation learning process of the transformer encoder. Finally, the joint semantics representation vectors are delivered to the experts (§3.2) and the manager (§3.3.2) for reasoning and management, respectively.

3.2 Mixture of Experts Module

In this part, a mixture of experts (MoE) module verifies the responses separately based on the joint semantics representation **H** extracted by (1).

We adopt three experts to focus on different part of the joint semantics representation, since the questions and the evidence can support the final prediction by interacting with the responses jointly or separately. Specifically, a question expert focuses more on the interaction between responses and questions, an evidence expert works for the interaction between responses and evidence, and a global expert takes responses, questions and evidence all into consideration.

However, different structures specially designed for the interactions among responses, questions and evidence would limit the generalization of the proposed framework to other datasets. Inspired by (Zhou et al., 2022), we implement each expert with the same general neural architecture but using different parameter learning strategies. Specifically, each expert is designed based on a stack of transformer encoding layers to obtain the final representation **h**. Then we feed **h** into a classifier to predict the probability of each label. The process above is formulated as follows:

$$\mathbf{h}_i = f_{Enc_i}(\mathbf{H}) \tag{2}$$

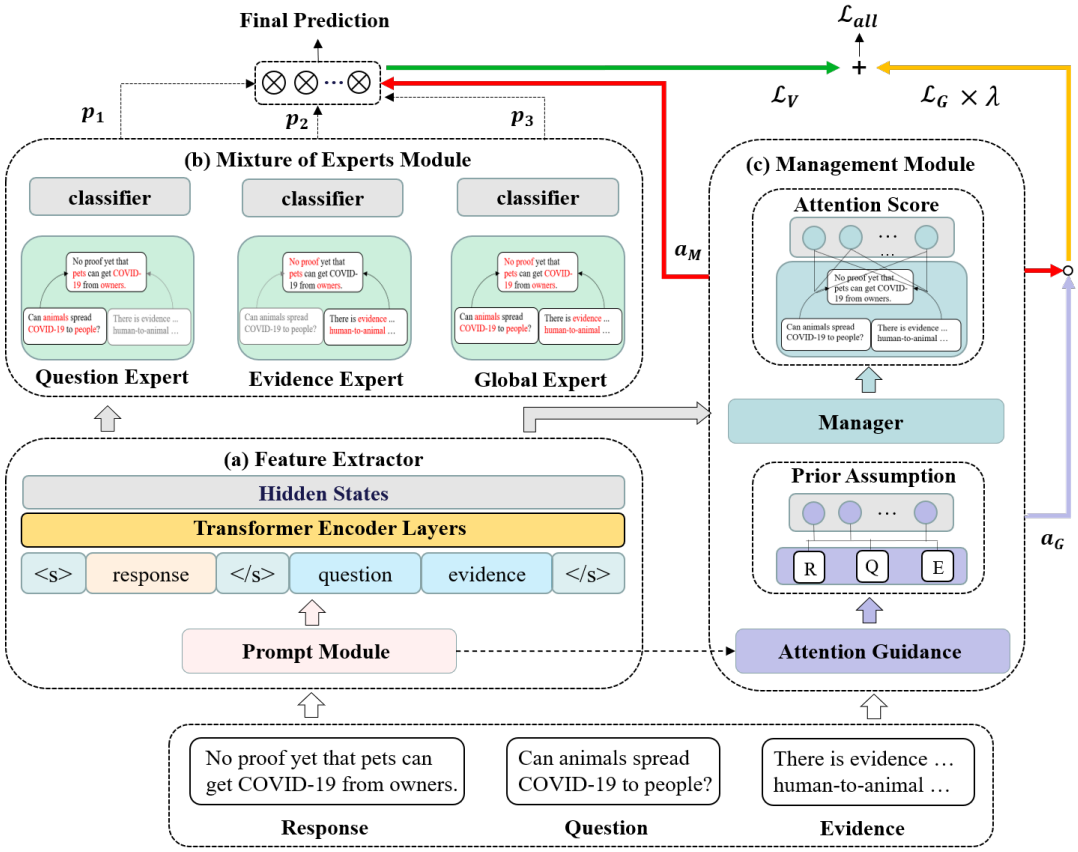


Figure 2: Model Architecture Overview of QaDialMoE. The model consists of three modules: (a) **Feature Extractor**: We propose a prompt module to generate *questions* by given *responses*, and directly concatenate the response, question and evidence embedding as a transformer-based encoder input. (b) **Mixture of Experts Module**: We use three experts to verify the responses separately against the same response-question-evidence joint semantics. (c) **Management Module**: We introduce an attention guidance module to generate prior assumption (R, Q and E mean response, question and evidence, respectively) and guide the manager assignment, then the manager summarizes the full output of experts as the final prediction.

$$\mathbf{p}_i = softmax(tanh(\mathbf{h}_i\mathbf{W}_1^i)\mathbf{W}_2^i) \quad (3)$$

Here, $f_{Enc_i}, i = 1, ..., n$ are n expert encoder networks and $\mathbf{h}_i \in \mathbb{R}^d$ refers the final representation vector encoded by the i^{th} expert, which implies different understanding to the relationship among responses, questions and evidence. The probabilities $\mathbf{p}_i$ is the prediction result from i^{th} expert, $\mathbf{W}_1^i$ and $\mathbf{W}_2^i$ are the trainable matrices of i^{th} expert's classifier, which projects $\mathbf{h}_i$ to the probabilities $\mathbf{p}_i$. $tanh$ and $softmax$ are activation functions.

3.3 Management Module

An attention guidance module is proposed to generate prior assumptions based on response-question-evidence pair and guide the manager. The manager is designed to guide experts' training and ensemble the results from all experts, which is implemented based on transformer (Vaswani et al., 2017) model.

3.3.1 Attention Guidance

Previous evidence-based fact verification models always incorrectly predict due to the word overlap issue. Since question-answering dialogue has both question and evidence parts to verify the response, an attention guidance module generates prior assumption that can represent the interactions among responses, questions and evidence, and guide the manager (§3.3.2) to focus more on questions or evidence based on their overlap degree with the response.

Specifically, the attention guidance module generates the prior assumption $\mathbf{a}_G$ based on the response-question-evidence pair (§3.1.1). In this paper, we consider three interactions, including response-question pair, response-evidence pair and response-question-evidence pair. We calculate the prior assumption $\mathbf{a}_G$ as follows:

1. Initialize a prior assumption with $\mathbf{z}_0 \in$

$\mathbb{R}^3$, which is empirically set as $\mathbf{z}_0 = ((\mathbf{z}_0)_0, (\mathbf{z}_0)_1, (\mathbf{z}_0)_2)^T = (0.2, 0.2, 0.6)^T$. The $(\mathbf{z}_0)_2$ represents the questions and the evidence interact with the responses jointly. It is always set higher than other values since we anticipate that this interaction can combine all the information efficiently.

2. Initialize a zero bias vector $\delta \in \mathbb{R}^3$ and calculate the response-question pair and response-evidence pair similarity scores s by TF-IDF. Then the similarity scores can be accumulated to the bias vector $\delta : \delta_i = a_i(1 - s_i)$[2], where δ_i is the i^{th} dimension of bias vector and s_i is the similarity score accumulated to δ_i (e.g., TF-IDF similarity score of response and question). $a_i \in (0, 0.4)$[3] is an incremental rate (set empirically) for the i^{th} dimension of bias vector δ.

3. Add the bias vector δ to the initialized assumption $\mathbf{z}_0$ and normalize to obtain the prior assumption: $\mathbf{a}_G = softmax(\mathbf{z}_0 + \delta)$.

The prior assumption $\mathbf{a}_G$ is used to teach the manager to assign scores reasonably against the attention scores $\mathbf{a}_M$ introduced in §3.3.2. Meanwhile, it can alleviate "imbalanced experts" phenomenon reported in previous studies (Eigen et al., 2013; Shazeer et al., 2017; Zhou et al., 2022). It means that the manager keeps assigning a close-to-1 attention score to one well-trained expert and a close-to-0 to other experts that are not trained efficiently.

3.3.2 Manager

We present a manager module to guide the training of experts. The manager encodes the joint semantics representation $\mathbf{H}$ and generates attention scores $\mathbf{a}_M$:

$$\mathbf{h}_M = f_{Enc_M}(\mathbf{H}) \tag{4}$$

$$\mathbf{a}_M = softmax(tanh(\mathbf{h}_M \mathbf{W}_1^M)\mathbf{W}_2^M) \tag{5}$$

where Enc_M is the encoder of manager module, $\mathbf{W}_1^M$ and $\mathbf{W}_2^M$ are trainable parameters. The manger has the same network architecture as experts, only the difference in the number of encoder layers and the dimension of output.

The attention score $\mathbf{a}_M$ and the prior assumption $\mathbf{a}_G$ are used to guide the experts' training and teach the manager to assign scores reasonably, which are implemented by specially designed losses introduced in §3.4.

[2]We want the manager paying more attention to the input part with few word overlap.

[3]The value of δ_i can not higher than 0.4 since the reason presented in the first step.

3.4 Loss

In this part, we develop two loss functions, i.e. verification loss $\mathcal{L}_V$ and guidance loss $\mathcal{L}_G$. The former one is a weighted sum of classification loss from each expert with the attention scores assigning to experts by manager as (6). The latter one measures the difference between the prior assumption and the attention scores, and guides the manager to assign reasonable attention scores to experts as given in (7). We jointly optimize our model by minimizing a weighted sum of these two terms: $\mathcal{L}_{all} = \mathcal{L}_V + \lambda\mathcal{L}_G$, where λ is a hyperparameter that controls the ratio of $\mathcal{L}_G$. The detail of these two loss functions are provided in subsequent paragraphs.

Verification Loss We calculate each expert's cross-entropy independently, which then is weighted by the attention scores $\mathbf{a}_M$ to sum up:

$$\mathcal{L}_V = \sum_{i=1}^{n_e} (\mathbf{a}_M)_i \cdot H_{CE}(\mathbf{p}_i, l) \tag{6}$$

where n_e is the number of experts, $(\mathbf{a}_M)_i$ is the i^{th} score assigned by manager for the i^{th} expert. The probabilities $\mathbf{p}_i$ is the prediction result from i^{th} expert, l is the ground true label of response and $H_{CE}(\cdot, \cdot)$ refers to the cross-entropy loss function.

Guidance Loss To alleviate the "imbalanced experts" phenomenon mentioned in §3.3.1, we develop another loss function $\mathcal{L}_G$, which calculates the logarithmic difference between the prior assumption $\mathbf{a}_G$ and the attention scores $\mathbf{a}_M$:

$$\mathcal{L}_G = D_{KL}(\mathbf{a}_G || \mathbf{a}_M) \tag{7}$$

where $D_{KL}(\cdot||\cdot)$ stands for the Kullback–Leibler divergence. By minimizing $\mathcal{L}_G$, the manager learns to assign reasonable attention scores to experts which means that the manager assigns each expert based on the interactions presented in §3.3.1. Besides, the training of experts become more balanced due to loss function $\mathcal{L}_G$.

4 Experiments

4.1 Dataset and Evaluation Metrics

Our model is evaluated on three benchmark datasets, i.e. HEALTHVER (Sarrouti et al., 2021),

FAVIQ (Park et al., 2022) and COLLOQUIAL (Kim et al., 2021). These datasets are introduced as follows and the statistic information is shown in Appendix A.

HEALTHVER HEALTHVER (Sarrouti et al., 2021) contains 14,330 real-word responses retrieved by a search engine for 80 popular questions about COVID-19. Each instance in HEALTHVER consists of a question, an evidence from scientific article and a response manually annotated as SUPPORT, REFUTE and NEUTRAL. Metrics as macro precision, macro recall, macro F1-score, and accuracy are used to evaluate the effectiveness of our model on HEALTHVER.

FAVIQ FAVIQ (Park et al., 2022) is a large-scale fact verification dataset constructed from information-seeking questions (Kwiatkowski et al., 2019) and their ambiguities (Min et al., 2020). The data consists of two sets (A and R), FAVIQ A set is obtained from ambiguous question-answering pairs while FAVIQ R set uses the reference answer and the incorrect prediction to generate responses. Most instances in FAVIQ include a question, an evidence from Wikipedia passages and a response annotated as SUPPORT and REFUTE. However, the questions for A test set is hidden since the A set is made from AmbigQA (Min et al., 2020), and there is a leaderboard[4] for the competition. For A set, we only use A dev set and do not generate questions for A test set. We use the accuracy as our evaluation metric.

COLLOQUIAL COLLOQUIAL (Kim et al., 2021) is constructed by transferring the styles of claims from FEVER (Thorne et al., 2018) into colloquialism. The data is challenging due to the characteristics of colloquial claims. Most question-answering dialogue based responses are colloquial style, our model can be easily generalized to this dataset since the prompt module can generate questions for the claims. Finally, each instance in COLLOQUIAL consists of a synthetic question, an evidence from Wikipedia passages and a response with label SUPPORTED, REFUTED or NEI. We use the label accuracy as our evaluation metric.

[4] https://nlp.cs.washington.edu/ambigqa/leaderboard.html

4.2 Implementtation Details

We download pre-trained models from huggingface[5]. For prompt module, we utilize 't5-small-squad2-question-generation', which is built based on SQuAD 2.0 dataset (Rajpurkar et al., 2018). We use 12 transformer encoder layers for the feature extractor and experts, and 2 for the manager, and we leverage RoBERTa-Large (Liu et al., 2019) to initialize parameters of the feature extractor and experts. For MoE module, we set $n_e = 3$ meaning three experts in our implementation of QaDialMoE. We apply Adam optimizer (Kingma and Ba, 2015) in training with learning rate 2e-5. The dimension of hidden states, maximum sequence length, batch size are 1024, 512, and 32, respectively. The a_i in §3.3.1 and the λ in §3.4 are set to 0.2 and 0.1, respectively. All codes are implemented with PyTorch (Paszke et al., 2019).

4.3 Baselines

We compared our proposed QaDialMoE model with different baselines on three benchmarks. (1) **HEALTHVER**: BERT (Devlin et al., 2019), SciBERT (Beltagy et al., 2019), BioBERT (Schneider et al., 2020) and T5 (Raffel et al., 2019). Note SciBERT and BioBERT are the variants of BERT. T5 is the prior state-of-the-art effectiveness. (2) **FAVIQ**: three different variants based on BART (Lewis et al., 2020), i.e. claim only BART, TF-IDF + BART and DPR + BART (Qu et al., 2021), FiD (Izacard and Grave, 2020) and FiD + **E**videntiality-guided **G**enerator (EG) (Asai et al., 2021). (3) **COLLOQUIAL**: BERT (Devlin et al., 2019), CorefBERT (Ye et al., 2020) with Kernel Graph Attention Network (KGAT) (Liu et al., 2020).

4.4 Results and Analysis

4.4.1 Main Results

Table 1, Table 2 and Table 3 summarize the experimental results of various models on HEALTHVER, FAVIQ and COLLOQUIAL, respectively.

HEALTHVER We evaluate the performance of our approach on HEALTHVER based on the questions and evidence in the original dataset. As shown in Table 1, QaDialMoE outperforms all the baselines by a large margin. On the test set of HEALTHVER, QaDialMoE reaches an accuracy of **84.26%**, achieving a new state-of-the-art on the

[5] https://huggingface.co/

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
QaDialMoE	**83.95**	**82.83**	**83.29**	**84.26**

Table 1: Comparative performance on HEALTHVER test set.

dataset, which is **3.57%** higher than the previous best one. Meanwhile, QaDialMoE outperforms T5 (the best method) with **3.13%**, **3.83%** and **3.69%** improvements in macro precision, macro recall, macro F1-score.

FAVIQ For evidence retrieval on FAVIQ, we use three ways to obtain evidence E: (1) DPR (Qu et al., 2021), (2) **e**videntiality-guided **g**enerator (**EG**) (Asai et al., 2021), and (3) the *positive evidence* (**PE**) in the original dataset. First, we use k passages as evidence($k = 3$), which are retrieved by a dual encoder based model DPR. Following Park et al. (2022), this baseline is jointly trained on the A set and the R set. Second, the generator uses a leave-one-out approach (Asai et al., 2021) to evaluate which evidence provide sufficient information. Third, the *positive evidence* is the top passage that contains the answer to the original question which is retrieved by TF-IDF (Park et al., 2022).

Table 2 presents the performance of various models on FAVIQ. For A set, we evaluate the performance of our approach with the evidence from above three ways. QaDialMoE also achieves a new state-of-the-art with an accuracy of **78.7%** by using the *positive evidence*. Note that QaDialMoE outperforms prior methods based on the same evidence, achieving significant improvements with **3.9%** (70.8% vs. 66.9%) and **5.3%** (74.9% vs. 69.6%) for DPR and EG, respectively. For R set, we evaluate our approach with DPR and the *positive evidence*. Since the baseline with DPR is jointly trained on the A set and the R set, where the R set mainly provides a source for data augmentation, QaDialMoE achieves comparable results on the R set but improves considerably on the A set. However, QaDialMoE with the *positive evidence* can reach remarkable performances with **86.1%** on the dev set and **86.0%** on the test set. In short, QaDialMoE achieves a new state-of-the-art result on the large-scale challenging fact verification benchmark FAVIQ.

Model	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**

Table 2: Fact verification accuracy on FAVIQ. We do not evaluate our model on FAVIQ A test due to the reason presented in §4.1.

Model	Document Retrieval +Evidence Selection	Label Accuracy
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**

Table 3: Fact verification label accuracy on COLLOQUIAL.

COLLOQUIAL We use the prompt module to generate questions for the whole test set in COLLOQUIAL. Meanwhile, a part of them is used for training and a part for validation. Some synthetic question Examples for COLLOQUIAL are shown in Appendix B. Table 3 shows the performance of various models on the test set of COLLOQUIAL, where QaDialMoE again obtains a new state-of-the-art label accuracy as **89.5%**. This improvement is also surprise to us since COLLOQUIAL does not have original questions while we generate synthetic questions by our prompt module. Colloquial claims tend to include filter words (e.g., *"yeah"*, *"you know"*), comments, or personal opinions which do not require a verification. However, our synthetic questions may help the model to focus more on what requires a verification, and ignore the above mentioned distractions.

4.4.2 Ablation Study

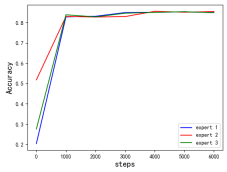
We further investigate the impact of question quality with an ablation study on FAVIQ A dev set. Specifically, the prompt module generates questions for both FAVIQ A train set and dev set and

Model	Accuracy
QaDialMoE + EG	74.9
- w/ synthetic questions	69.7 (**-5.2%**)
QaDialMoE + PE	78.7
- w/ synthetic questions	75.9 (**-2.8%**)

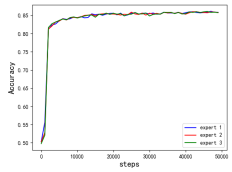
Table 4: Ablation study on FAVIQ A dev set. It shows the results of using synthetic questions rather than original questions.

Models	P	R	F1	Acc.
QaDialMoE	**83.95**	**82.83**	**83.29**	**84.26**
- w/o $\mathcal{L}_G$	82.39	82.01	82.18	83.04
- w/ fixed a_G	83.06	80.96	81.68	82.94

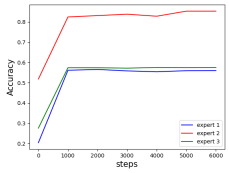
Table 5: Ablation study on HEALTHVER test set. It shows the results of training without the guidance loss $\mathcal{L}_G$ and with a fixed prior assumption a_G.



(a) Trained on HEALTHVER



(b) Trained on FAVIQ



(c) Trained on HEALTHVER without the Guidance Loss

Figure 3: The differentiation of experts. We show the model trained on HEALTHVER and FAVIQ R set with the *positive evidence*, and $n_e = 3$.

then we train and evaluate our model with the synthetic questions rather than original questions. Some synthetic questions as examples for FAVIQ A dev set are shown in Appendix B. The results of ablation study are presented in Table 4. It is clear that our QaDialMoE model has significant drops by **5.2%** with the *evidentiality-guided generator* and **2.8%** with the *positive evidence*. The prompt module can provide high quality questions, however, it is far less effective than using the original questions. Note that the effectiveness of using *evidentiality-guided generator* drops more noticeably than *positive evidence*, which verifies that high-quality questions and evidence play an important role equally.

Effects of the guidance of the manager. We have two ablative groups as shown in Table 5:

w/o $\mathcal{L}_G$: We conduct an ablation study on HEALTHVER dataset without the guidance loss $\mathcal{L}_G$. As presented in Table 5, QaDialMoE model has drops by **1.1%** (83.29% vs. 82.01%) and **1.2%** (84.26% vs. 83.04%) in macro F1-score and accuracy while training without the guidance loss. Meanwhile, the "imbalanced experts" phenomenon (§3.3.2) will be discussed more detailed in Sec. 4.4.3.

w/ fixed a_G: We initialize the prior assumption a_G with the same weights for each parameter and do not use the inverse TF-IDF similarity to correct. It means that each expert is equally important. QaDialMoE model also has a significant drop with this setting of prior assumption. Besides, we find that this model is even less effective than the model without the guidance loss. It means negative settings for the prior assumption may lead to inferior model results. The Full model with inverse TF-IDF similarity correcting the prior assumption is a positive setting. As aforementioned, since the word overlap issue, we guide the manager to pay more attention to the input part with few word overlap, which is proven to have a good performance.

4.4.3 Analyzing Experts Differentiation

To further understand the effectiveness of the proposed framework, we investigate the differentiation of experts, which means that the model can achieve balanced training across experts based on the prior assumption from the attention guidance. As shown in Figure 3a and 3b, each expert is well-trained and the "imbalanced experts" phenomenon does not occur. Note that there is no unique expert always outperforms others, which illustrates that experts behave independently due to the various interactions among responses, questions and evidence. However, as shown in Figure 3c, once training is performed without the guidance loss, there is only one well-trained expert, and the performance of the other two experts stay around 50% while training steps increasing. It seems the model degenerates to the point where only one RoBERTa-Large (Liu et al., 2019) is working, which is a simpler model, but far less effective than the Full model.

5 Related Work

Fact Verification To mitigate the spread of false information online, a fact verification task has gained widespread attention recently. Previous

studies on the fact verification are mainly based on pieces of evidence from Wikipedia articles (Thorne et al., 2018; Hanselowski et al., 2018; Yoneda et al., 2018; Thorne et al., 2019; Nie et al., 2019; Liu et al., 2020). Since the proposal of the TABFACT (Chen et al., 2020), a large dataset for table-based fact verification, studies against semi-structured evidence attach much attention (Zhong et al., 2020; Shi et al., 2020; Yang et al., 2020; Eisenschlos et al., 2020; Shi et al., 2021; Liu et al., 2021). However, fact verification in a question-answering dialogue is still an underexplored area. Gupta et al. (2021) explored fact verification for the dialogue context, curated by converting grounding dialogues from the Wizard-of-Wikipedia (Dinan et al., 2019b) dataset. Meanwhile, several works have used question-answering dialogue data to construct fact verification benchmarks (Demszky et al., 2018; Jiang et al., 2020; Pan et al., 2021; Chen et al., 2021; Sarrouti et al., 2021; Park et al., 2022). Different from previous works, we formulate the question-answering dialogue based fact verification task, which focuses on various interactions among responses, questions and evidence that support the validate process.

Mixture of Experts Mixture of experts is an ensemble learning method that first introduced by Jacobs et al. (1991), which is used to divide the problem space into homogeneous regions (Baldacchino et al., 2016). Specifically, it first decomposes a task into sub-tasks and then trains an expert model on each sub-task, a gating model is applied to learn which expert is competent and combine the predictions. Mixture of experts has been applied in various fields, such as dialogue systems (Le et al., 2016b), content recommendation (Ma et al., 2018; Zhu et al., 2020) and image classification (Wang et al., 2020; Riquelme et al., 2021). Zhou et al. (2022) develop a mixture-of-experts framework for table-based fact verification, where each expert is used to deal with different types of reasoning. In this paper, we leverage a mixture-of-experts module to recognize and execute various interactions among responses, questions and evidence, which we formulate as the question-answering dialogue based fact verification task.

6 Conclusion

In this paper, we present QaDialMoE, a new method for fact verification in question-answering dialogue that exploits the mixture of experts to focus on various interactions among responses, questions and evidence. We also generate synthetic questions with a prompt module to make our approach more generalizable. A manager with an attention guidance module is applied to guide the training of experts and assign a reasonable attention score to each expert. Experimental results on three datasets HEALTHVER, FAVIQ and COLLOQUIAL demonstrate that QaDialMoE outperforms previous approaches by a large margin and achieves new state-of-the-art results on all of them. The ablation studies and analysis further indicate that questions and evidence play an equal important role in our proposed framework. We hope our work can facilitate fact verification in a question-answering dialogue domain, and open the way to efficiently exploit questions and evidence in the verification process.

Limitations

The first limitation of our approach is the quality of synthetic question. As mentioned above, we employ the pre-tained model T5 to generate questions. It works well and can provide high quality questions, but is far less effective than using the original questions. In practice, when the quality of both questions and evidence are low, the performance of model will drop significantly. The second limitation is that the task of verifying responses in common dialogue cannot benefit from our proposed framework. We have tried to apply QaDialMoE in DialFact (Gupta et al., 2021), a benchmark for fact verification in dialogue, where questions in the input convert to the dialogue context. However, QaDialMoE does not show a significant advantage over the best method. We attribute this to two factors: first, the common dialogue based fact verification requires more sophisticated interactions among responses, dialogue contexts and evidence, since the common dialogue are more informal than question-answering dialogue; second, the benchmark consists of multi-turn dialogue while the final response needs to be validated. Rather than that simply replacing questions with dialogue contexts as part of the input, we may need to model the relationships between dialogue contexts.

Acknowledgements

The authors thank all the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development of China (No. 2021YFB3100600),

Youth Innovation Promotion Association of CAS (No. 2021153), and Strategic Priority Research Program of Chinese Academy of Sciences (No. XDC02040400).

References

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

Tara Baldacchino, Elizabeth J Cross, Keith Worden, and Jennifer Rowson. 2016. Variational bayesian mixture of experts models and sensitivity analysis for nonlinear dynamical systems. *Mechanical Systems and Signal Processing*, 66:178–200.

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

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

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. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational journalism: A call to arms to database researchers. In *CIDR*.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019a. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 281–296, Online. Association for Computational Linguistics.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 166–175.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2021. A survey on automated fact-checking. *arXiv preprint arXiv:2108.11896*.

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

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129.

Gautier Izacard and Edouard Grave. 2020. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

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*.

Phong Le, Marc Dymetman, and Jean-Michel Renders. 2016a. LSTM-based mixture-of-experts for knowledge-aware dialogues. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 94–99, Berlin, Germany. Association for Computational Linguistics.

Phong Le, Marc Dymetman, and Jean-Michel Renders. 2016b. Lstm-based mixture-of-experts for knowledge-aware dialogues. In *1st Workshop on Representation Learning for NLP*, pages 94–99. ACL Anthology.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Qian Liu, Bei Chen, Jiaqi Guo, Zeqi Lin, and Jianguang 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*.

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

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1930–1939.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Salman Bin Naeem, Rubina Bhatti, and Aqsa Khan. 2021. An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal*, 38(2):143–149.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6859–6866. AAAI Press.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. 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 476–483, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. FaVIQ: Fact verification from information seeking questions. In *ACL*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

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.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, 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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Julien Knafou, Lucas Emanuel Silva e Oliveira, Jenny Copara, Yohan Bonescki Gumiel, Lucas Ferro Antunes de Oliveira, Emerson Cabrera Paraiso, Douglas Teodoro, and Cláudia Maria Cabral Moro Barra. 2020. BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 65–72, Online. Association for Computational Linguistics.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Qi Shi, Yu Zhang, Qingyu Yin, and Ting Liu. 2020. Learn to combine linguistic and symbolic information for table-based fact verification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5335–5346, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qi Shi, Yu Zhang, Qingyu Yin, and Ting Liu. 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, Online and Punta Cana, Dominican Republic. 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 *NAACL-HLT*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The second fact extraction and verification (fever2. 0) shared task. *EMNLP 2019*, page 1.

Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. 2019. Do sentence interactions matter? leveraging sentence level representations for fake news classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 134–139, Hong Kong. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. 2020. Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence*, pages 552–562. PMLR.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Xiaoyu Yang, Feng Nie, Yufei Feng, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2020. Program enhanced fact verification with verbalization and graph attention network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7810–7825, Online. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. *arXiv preprint arXiv:2004.06870*.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. UCL machine reading group: Four factor framework for fact finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium. Association for Computational Linguistics.

Wanjun Zhong, Duyu Tang, Zhangyin Feng, Nan Duan, Ming Zhou, Ming Gong, Linjun Shou, Daxin Jiang, Jiahai Wang, and Jian Yin. 2020. LogicalFactChecker: Leveraging logical operations for fact checking with graph module network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6065, Online. Association for Computational Linguistics.

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*.

Ziwei Zhu, Shahin Sefati, Parsa Saadatpanah, and James Caverlee. 2020. Recommendation for new users and new items via randomized training and mixture-of-experts transformation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1121–1130.

A Statistics of Experimental Datasets

Table 6 shows the statistics of HEALTHVER (Sarrouti et al., 2021), a dataset for fact verification of health-related real world responses.The whole dataset is split into three subsets for training, validation and testing by claims and thus have a balanced dataset class-wise.

Table 7 shows the statistics of FAVIQ (Park et al., 2022), a large-scale challenging fact verification dataset, which consists of 188k claims. FAVIQ-*A* is created from ambiguous questions, while FAVIQ-*R* includes claims from regular question-answering dialogues.

Table 8 shows the statistics of COLLOQUIAL (Kim et al., 2021), which transfers the claims from FEVER (Thorne et al., 2018) into a colloquial style. Each claim in COLLOQUIAL has three more words on average than in FEVER. We generate questions for the whole test set in COLLOQUIAL while a part for training and a part for validation.

Set	Supports	Refutes	Neutral	Total
Training	3,782	2,411	4,397	10,590
Validation	533	391	993	1,917
Test	671	425	727	1,823
Total	4,986	3,227	6,117	14,330

Table 6: Statistics of HEALTHVER dataset and the claim labels distribution.

B Synthetic Question Examples

COLLOQUIAL We present some examples of synthetic question generated by our prompt module for COLLOQUIAL:
Example 1:

- FEVER (REFUTES): Barack Obama will forgo a presidential library in favor of a presidential science museum.
- Colloquial claim: Oh yeah. Obama decided to forgo building a presidential library in favor of building a presidential science museum.
- Synthetic question: What did Obama decide to forgo?

Example 2:

- FEVER (REFUTES): Transformers: Revenge of the Fallen grossed a total of 836.4 million dollars worldwide.
- Colloquial claim: Yes they were, Transformers Revenge of the Fallen grossed 836.4 million dollars worldwide.
- Synthetic question: How many dollars did Transformers Revenge of the Fallen grosse?

Example 3:

- FEVER (SUPPORTS): Yung Rich Nation was produced by Zaytoven.
- Colloquial claim: I remember the song "Yung Rich Nation". It was produced by Zaytoven.

		Total	Support	Refute
Train	A	17,008	8,504	8,504
	R	140,977	70,131	70,846
Dev	A	4,260	2,130	2,130
	R	15,566	7,739	7,827
Test	A	4,688	2,344	2,344
	R	5,877	2,922	2,955

Table 7: Statistics of FAVIQ dataset, consisting of *A* set and *R* set.

	Train	Valid	Test	Words/Claim
FEVER	145.4K	10K	10K	8.2
COLLOQUIAL	410.0K	25.9K	8.4K	11.1
OURS	126.5K	8.5K	8.4K	-

Table 8: Statistics of COLLOQUIAL dataset compared to FEVER (Thorne et al., 2018) and the statistics of our generated questions.

- Synthetic question: What song was produced by Zaytoven?

Example 4:

- FEVER (SUPPORTS): Henry VIII of England had a war with the Holy Roman Emperor Charles V.
- Colloquial claim: Yep! Henry VIII, starting the war with Charles V.
- Synthetic question: Who started the war with Charles V?

Example 5:

- FEVER (SUPPORTS): An Emmy award was won by Mad Men.
- Colloquial claim: Yes, it is! Mad Men actually won an Emmy award.
- Synthetic question: What award did Mad Men win?

FAVIQ We generate questions for FAVIQ A dev set to investigate the impact of question quality with an ablation study. Table 9 shows the automatic evaluation results of the quality of synthetic question. Some examples of synthetic question are presented as follows:
Example 1:

- Response (REFUTES): tyrod taylor, ej manuel, matt cassel was the starting quarterback for the buffalo bills in 2016.

BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE_L
0.46	0.40	0.36	0.33	0.48

Table 9: Automatic evaluation results of the question quality for FAVIQ A dev set by BLEU 1–4 (Papineni et al., 2002) and ROUGE_L (Lin, 2004)

- Original question: who's the starting quarterback for the buffalo bills in 2016?
- Synthetic question: what quarterback was the starting quarterback for the buffalo bills?

Example 2:

- Response (REFUTES): the upper school of the minnehana academy is located at 4200 west river parkway, minneapolis, minnesota, 55406 in minneapolis.
- Original question: where is the upper school of the minnehana academy in minneapolis?
- Synthetic question: what is the upper school of the minnehana academy located at?

Example 3:

- Response (REFUTES): the new independence day came out in 1996 throughout the united states on june 24, 2016.
- Original question: when does the new independence day come out in 1996 throughout the united states?
- Synthetic question: when did the new independence day come out?

Example 4:

- Response (SUPPORTS): 11 players in one team can play on the field for american football.
- Original question: how many players in one team can play on the field for american football?
- Synthetic question: how many players in one team can play on the field for american football?

Example 5:

- Response (SUPPORTS): melinda o. fee played jill abbott on the young and restless on 1984.
- Original question: who played jill abbott on the young and restless on 1984?
- Synthetic question: who played jill abbott on 1984?