"It takes two to tango": A Combination of Closed-Domain and Open-Domain Few-Shot Prompting for Claim Verification

Anonymous ACL submission

Abstract

The widespread use of social media platforms has resulted in the swift dissemination of misinformation and fake news, creating a critical need for the development of computational models for automated fact-checking. Existing work on claim verification mainly relies on supervised learning from manually annotated claim-evidence pairs, which is resource-intensive and prone to biases, limiting their generalization across domains. To address this gap, we investigate zero-shot domain adaptation for claim verification, where no labeled training data is available for the target domain. We propose a hybrid approach that combines utilizing labeled training data from a source domain via in-context learning, along with topically relevant contexts from target document collections such as Wikipedia by means of RAG. We conduct experiments to evaluate zero-shot domain adaptation of claim verification for three target domains, namely climate change, scientific publications, and COVID-19 with the training set of the FEVER dataset as the source domain. We find that our proposed approach outperforms supervised models for domain adaptation, several LLM prompting-based models including zero-shot, and few-shot prompting from the source domain, and an RAG-based approach over a target collection of Wikipedia.

1 Introduction

While on one hand, the social media provide platforms for individuals to access, contribute, and disseminate information, on the other hand, they also act as breeding grounds for rapid and widespread transmission of misinformation and fake news (Kumar et al., 2016; Chen et al., 2015). This has necessitated development of computational models of 'claim verification' or 'fact checking' with capabilities of automatically estimating the truthfulness or falsity of claims (Schuster et al., 2019, 2021; Jiang et al., 2021) by retrieving evidences for or against them (Asai et al., 2023).

The advent of large language models (LLMs) (Touvron et al., 2023; Wang and Komatsuzaki, 2022) has further aggravated the situation of fake news production at scale, because it is mostly straightforward to programmatically generating misinformation via LLMs with the help of suitably crafted adversarial prompts (Zou et al., 2023). The topically coherent and fluent nature of an LLM-generated text (Liu et al., 2021b) potentially makes it even harder to detect any injected misinformation (Parry et al., 2024). Moreover, LLMs, due to the inherent stochastic nature of their generative process, are reported to inadvertently generate factually incorrect content - a phenomenon commonly referred as hallucinations (Zhang et al., 2023); this LLM-hallucinated content, when published without fact checking on online platforms, further contributes to the volume of misinformation.

Standard computational approaches for claim verification involve pairwise supervised learning from claim-evidence pairs (Gururangan et al., 2018), which means that training these models requires manual annotation of relevant evidence for or against each claim (Poliak et al., 2018). The standard practice to obtain a test collection of manual annotated claim-evidence pairs is as follows: given a claim, a top-retrieved set of text segments (e.g., with BM25) is obtained from an indexed collection, such as Wikipedia, and subsequently the relevance of these segments is assessed manually as evidences to support or refute the claim (Thorne et al., 2018a). Not only does it cost time, effort, and financial resources to compile such a dataset large enough to train supervised models, but the dataset constructed this way is also likely to exhibit pooling biases (Buckley et al., 2007; Gao et al., 2022) due to a small number of top-documents used to decide the truth of a claim.

Due to an inherent anchoring effect of relating a claim only to a small subset of evidences means that supervised models trained on such claim-evidence pairs (Stammbach and Neumann, 2019a; Krishna et al., 2022) are likely to be exhibiting biases and thus generalize poorly to a different domain (Pan et al., 2023; Talmor and Berant, 2019; Hardalov et al., 2021). While investigation of out-of-domain (OOD) generalization of predictive models has been extensively carried out for a range of diverse tasks, such as information retrieval (Thakur et al., 2021; Kim et al., 2023), named entity recognition (NER) (Long et al., 2022), question answering (Labruna et al., 2024), speech emotion recognition (Lashkarashvili et al., 2024), several prediction tasks in the clinical domain, such as predicting the treatment, diagnosis, in-hospital mortality etc. (Gema et al., 2024), to the best of our knowledge there exists no work that has explored OOD claim verification.

To bridge this research gap, in this paper we explore the task of **zero-shot domain adaptation for claim verification**, i.e., we assume that labeled training data exists only for a source domain, and that the **target domain is devoid of any training data**. The core hypothesis underlying our work is that a parametric memory acquired from a source domain may not yield effective results for a target domain, in which case non-parametric memory, e.g., via the use of in-context learning (ICL) (Izacard et al., 2023; Liu et al., 2022; Lu et al., 2022), may help improve OOD effectiveness.

Our Contributions. Following is a list of contributions of this paper.

- To the best of our knowledge, this work of ours is the first to investigate zero-shot out-of-domain claim verification via in-context learning (ICL) and retrieval augmented generation (RAG).
- We propose to combine the two sources of information - one from a training set of a (source) domain - which is different from the target one, and the other from an external collection of documents (specifically Wikipedia), to improve the effectiveness of claim verification.
- An extensive set of experiments on three different claim verification tasks on climate, scientific publications, and the Covid disease, with zero-shot OOD transfer from FEVER (Thorne et al., 2018b) (generic domain labeled examples of claims and evidences) shows the efficacy of our proposed approach.

We also make our source code[1] available for research purposes.

2 Related Work

In-Context Learning. The effectiveness of pre-trained language models (PLMs) for few-shot learning is suboptimal due to the gap between pre-training and downstream tasks. GPT-3 introduced prompt tuning, using natural language prompts and demonstrations (Brown et al., 2020). Recently, large language models like GPT-3.5 have excelled in various tasks (Wei et al., 2022; Zhou et al., 2022). In-context learning (ICL) provides an alternative by conditioning on demonstration examples without training (Brown et al., 2020), enabling tasks like fact verification through Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Zhang and Gao, 2023).

Fact Checking. Fact-checking methodologies often verify trustworthy sources, retrieve evidence, and assess the veracity of claims. Recent research on Fact Extraction and Verification (FEVER) includes supervised approaches using pre-trained models (Stammbach and Neumann, 2019b; Krishna et al., 2022), multitask learning (Hidey and Diab, 2018), and retrieval models (Lewis et al., 2020). Some studies use Graph Neural Networks for verification (Zhao et al., 2020; Zhong et al., 2019). There is also a focus on using the web for evidence retrieval and unsupervised methods to reduce annotation costs (Subramanian and Lee, 2020; Stammbach, 2021).

Our work differentiates from existing literature by combining closed-domain in-context learning (CICL) with open-domain in-context learning (OICL), leveraging both annotated examples and external contextual information, to achieve better zero-shot domain adaptation in fact verification tasks.

3 LLM-based Claim Verification

The task of fact verification involves assessing the truthfulness or falsity of a claim by retrieving contexts for or against it (Thorne et al., 2018b). Our objective is to analyze the impact of domain adaptation on the downstream fact verification task, specifically aiming to adapt the knowledge acquired from labeled examples of claim evidence pairs from the source domain towards an effective generalization in the target domain.

In this section, we first provide a brief overview of existing supervised approaches for the claim verification task. This is followed by an overview of

[1] https://anonymous.4open.science/r/Misture_of_Experts-DC6A/

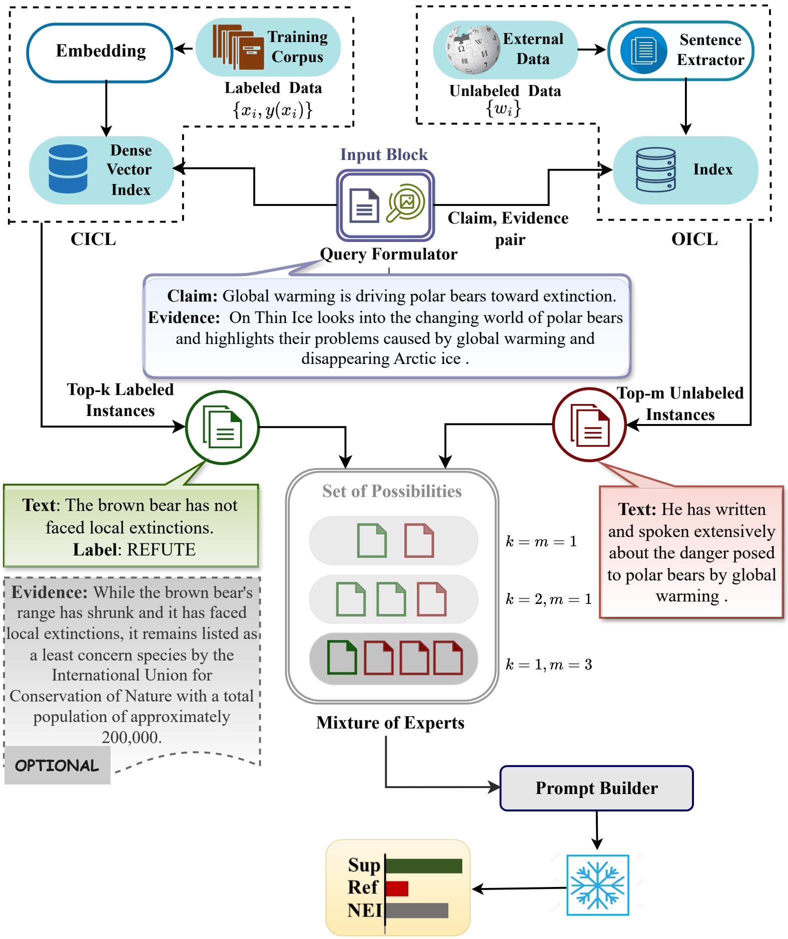


Figure 1: Schematic diagram of our proposed framework for downstream claim verification task. After a given claim-evidence pair is passed through a query formulator, the top-k labeled instances from the training corpus and the top-m documents retrieved from the external corpus are combined to determine the validity of that particular claim. Here, both k and m depend on two parameters - α – the relative proportion of unlabelled data to labelled one), and M – the maximum number of data units (either labelled or unlabelled) to consider (see Equation 4). The diagram shows different possible combinations with which our model can be instantiated as the different sets of values for the (k, m) pair.

in-context learning (ICL) based approach of leveraging labeled examples, and also that of retrieval augmented generation (RAG), which utilizes additional context (unlabelled data) from collections, such as the Wikipedia. We then describe how we combine the two approaches of ICL and RAG – we call the former closed-domain ICL (CICL), and the latter open-domain ICL (OICL) – for the task of claim verification.

3.1 Background

Supervised approach. Given a claim $\mathbf{x}$ (bag-of-words representation of text, or its dense vector embedding, e.g., as obtained by an encoder, such as BERT (Devlin et al., 2019)), and a collection $\mathcal{C}$ of documents, the task of (closed-domain) claim verification is that of a 3-way classification one, i.e., the task requires predicting a label $y(\mathbf{x}) \in \{0, 1, 2\}$, where the labels map to the three possibilities of whether the claim is 'support' or 'refute' by relevant evidences from the collection, or there is not enough evidence in the collection to arrive at one of these two decisions about the claim. Retrieving a set of topically relevant candidate evidences with a query formulated from the claim is an intermediate task for claim verification. More formally, the prediction function takes the form

$$\phi : (\mathbf{x}, \mathcal{R}_m(\mathbf{x})) \mapsto \Delta_3, \qquad (1)$$

where each $e \in \mathcal{R}_m(\mathbf{x})$ is a set of m candidate evidences retrieved from the collection, and Δ_3 denotes the posterior probability distribution simplex of the three output classes.

To learn this function ϕ, in a supervised manner, existing supervised approaches either make use of an already available training set of manually labeled ground-truth claim-evidence pairs (Thorne et al., 2018b), or make use only of the claim-level annotations associating them to the top-retrieved evidences to infer weak labels for training (Atanasova et al., 2022).

Closed-Domain In-context Learning (CICL). Since supervised models require a large quantity of data for effective training, and is also likely to not generalize well to specific domains, researchers have started to explore the potential benefits of the semantic capabilities of large language models (LLMs) for this task of claim verification. Unlike learning a parameterized representation of the function $\phi(\mathbf{x}, \mathcal{R}_m(\mathbf{x}))$, an ICL-based workflow retrieves a small number of examples that are similar to the current claim from a training set $\mathcal{T}$ of claim-evidence pairs eventually including these as a part of an input prompt to an LLM (Liu et al., 2021a; Agrawal et al., 2022; Huang et al., 2022). Formally,

$$\phi_{\text{LLM}}(\mathbf{x}, \mathcal{N}_k(\mathbf{x})) \mapsto \text{<MASK>}, \tag{2}$$

where <MASK> is the output generated by an LLM indicating the name of the output class (i.e., one of 'support', 'refute', or 'not enough information'), and $\mathcal{N}_k(\mathbf{x}) \subset \mathcal{T}$ is the set of claim-evidence pairs from the training set $\mathcal{T}$ that are most similar (in terms of lexical or semantic similarity) to the input claim $\mathbf{x}$.

Open-Domain In-context Learning (OICL). Different from CICL, where similar examples are prompted to an LLM, in open-domain ICL, an additional context in the form of candidate evidences retrieved from the collection $\mathcal{C}$ is fed as a part of the input prompt to an LLM. This means that OICL does not require any training set $\mathcal{T}$ of labeled examples. Stated explicitly,

$$\phi_{\text{LLM}}(\mathbf{x}, \mathcal{R}_m(\mathbf{x})) \mapsto \text{<MASK>}. \tag{3}$$

3.2 Proposed Methodology

We now describe our proposed methodology which utilizes the best of both worlds by combining both the labeled data via CICL (Equation 2) and unlabeled data in the form of potentially relevant candidate evidences via OICL (Equation 3). Both the approaches use individual hyper-parameters to control the quantity of information fed as input to an LLM prompt, i.e., k to control the number of examples in few-shot prompting, vs. m to control the number of candidate evidences. To allow a general combination of the two approaches in varying proportions, we define a hyper-parameter as $\alpha \in [0, 1]$. The combined methodology then uses an $\alpha : 1-\alpha$ proportion of data for OICL and CICL. More formally,

$$\phi_{\text{LLM}}(\mathbf{x}, \mathcal{N}_{\lfloor (1-\alpha)M \rfloor}(\mathbf{x}), \mathcal{R}_{\lfloor \alpha M \rfloor}(\mathbf{x})) \mapsto \text{<MASK>}, \tag{4}$$

where $\lfloor x \rfloor$ denotes the floor function, i.e., the largest integer not greater than x, and M is an upper bound on the number of sentences over which the relative proportions are defined. Equation 4 implies that instead of being functions of k (CICL) and m (OICL), the predictor uses variable contributions from both, as parameterized by α and M. To make Equation 4, consistent with Equations 2 and 3, call the number of example sentences for CICL and OICL, k and m, respectively, with $k \equiv \lfloor (1-\alpha)M \rfloor$ and $m \equiv \lfloor \alpha M \rfloor$. We call our methodology **Mixture of Experts** (**MoE**). The prompt used in the MoE method along with an example claim instance is shown in Figure 2.

4 Evaluation

4.1 Experiment Setup

We hypothesize that our proposed approach of MoE-based ICL is particularly suitable for out-of-domain OOD generalization tasks. As such, we conduct experiments to evaluate the quality of zero-shot transfer from a source domain to a target domain, i.e., the target domain is devoid of any training data. To this end, we compare our approach with standard non-parametric approaches of LLM-based prompting (0-shot and few-shot), and also with supervised models involving low rank adaptation for domain transfer.

For our experiments, as the target domain we consider the following three datasets:

- Climate-FEVER (Diggelmann et al., 2020; Thakur et al., 2021): dataset comprised of claims and evidences related to the climate change;
- SciFact (Wadden et al., 2020): dataset constituting scientific claims;
- COVID (Wang et al., 2023): a dataset of correct and incorrect facts related to the Covid pandemic.

We only use the test splits of the above datasets for

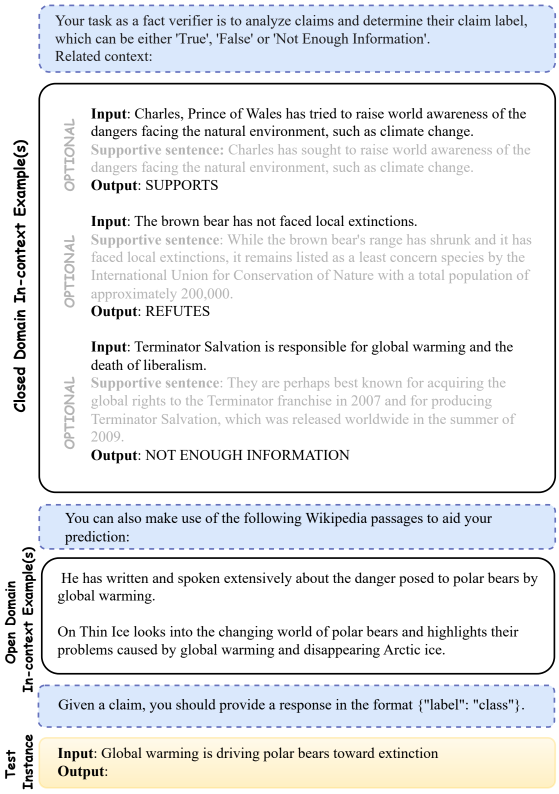


Figure 2: An illustration of the prompt structure used in our proposed approach of Mixture-of-Experts (MoE)-based ICL. In this example, M, the total number of example or context sources over which the relative proportions are defined (see Equation 4), is 3, and $\alpha = 2/3$. This means that $k = 1$ (in our setting, 1 example for each class), and $m = 2$. The blue segments refer to the instructions, the white segments show the examples being included (either retrieved from a training set in CICL, or from the Wikipedia), and the yellow segment shows a sample claim (the current test instance).

evaluation (since Climate-FEVER has no train:test split, we use the entire set for evaluation).

As the dataset corresponding to the source domain, we use the train split of FEVER (Thorne et al., 2018b) as the source of labeled data examples to be used in the supervised and the in-context learning approaches. For all these datasets, the target collection, i.e., the collection of documents used to retrieve potentially relevant evidences, is the Wikipedia dump of 2018. Table 1 summarizes these datasets.

3-way vs. 2-way classification. Researchers usually treat the claim verification task as a 3-way classification problem (Pan et al., 2023), where the labels for given claim evidence pair are either 'Support', 'Refute', or 'Not Enough Info'; we call this standard setup by the name '**SRN**'.

Dataset	Usage	Labels (S:R:N)	#Claims
FEVER	Train	52:22:26	145,327
Climate-FEVER		47:18:34	1,381
SciFact	Test	41:21:37	300
COVID-C		32:36:33	180

Table 1: Fact verification datasets used in our experiments for zero-shot domain adaptation. The three classes are abbreviated as 'S' (support), 'R' (refute), and 'N' (not enough information), and their proportions are reported as percentages.

In contrast to the 3-way (SRN) setup, some authors, e.g., Pan et al. (2023); Jiang et al. (2020); Saakyan et al. (2021) do not consider the claims with label 'NEI' for training or evaluation. This makes the experiment setup less ambiguous thus likely leading to more conclusive outcomes (Poliak et al., 2018). We name this setup '**SR**'.

LLM Details. We conduct all our experiments using the LLM LLAMA-2.0[2] (70B) model (Touvron et al., 2023). This choice follows a set of initial experiments with various LLMs, exploring different options for ϕ_{LLM} in Equation 2. LLAMA-2.0 consistently outperformed the other models for this task. All transformer models in our experiments use the HuggingFace API. For the LLM-based setup, we utilize the vLLM[3] (Kwon et al., 2023) library to apply k-v cache optimization, enhancing computation speed. For finetuning the supervised approaches in our experiments (specifically, RoBERTa[4] and LLaMA-LoRA[5]; more details in Section 4.2), we use source training dataset for 10 epochs using AdamW (Loshchilov and Hutter, 2019) as the optimizer with a learning rate of $5e-5$; the training batch size used was 8. Additionally, we apply a parameter efficient finetuning (PEFT) (Xu et al., 2023) based strategy - specifically, a Low Rank Adaptation (LORA) (Hu et al., 2021) technique for tuning LLAMA.

4.2 Methods Investigated

We compare our proposed methodologies with the following baselines.

Non-parametric baselines. These methods do not involve any parametric training on the labeled

[2] `https://huggingface.co/TheBloke/Llama-2-70B-Chat-AWQ`

[3] `https://github.com/vllm-project/vllm.git`

[4] `FacebookAI/roberta-base`

[5] `huggyllama/llama-7b`

	Experiment	SR						SRN					
		Climate Fever		SCIFACT		Covid C		Climate Fever		SCIFACT		Covid C	
	Setup	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Supervised	RoBERTa	0.734	0.638	0.697	0.599	0.694	0.686	0.370	0.229	0.383	0.237	0.367	0.324
	LLaMA-LoRA	0.666	0.591	0.729	0.665	0.587	0.602	0.448	0.357	0.450	0.348	0.401	0.340
Unsupervised	0-shot	0.772	0.731	0.745	0.664	0.686	0.680	0.520	0.429	0.470	0.359	0.444	0.365
	CICL-E	0.469	0.469	0.649	0.622	0.645	0.643	0.442	0.343	0.480	0.378	0.406	0.389
	CICL	**0.798**	0.741	0.676	0.564	0.645	0.635	0.519	0.456	0.440	0.352	0.422	0.398
	OICL	0.749	0.718	0.819	0.809	**0.835**	**0.833**	0.514	0.450	0.557	0.473	**0.551**	0.449
	MoE-E (Ours)	0.785	0.737	0.825	0.782	0.802	0.801	0.526	0.478	0.520	0.461	0.501	**0.471**
	MoE(Ours)	0.783	**0.742**	**0.840**	**0.810**	0.826	0.826	**0.542**	**0.509**	**0.583**	**0.550**	0.502	0.454

Table 2: Macro-F1 and overall accuracy of the 3-way and the 2-way evaluation of fact verification. For these results, the MoE approach uses $\alpha = 0.5$ (Equation 4).

examples from the FEVER dataset.

- **0-shot**: Labruna et al. (2024); Kojima et al. (2022); Li et al. (2023) investigate the efficacy of leveraging the pretrained knowledge of LLMs on various downstream tasks. We follow a similar pathway by prompting the LLM for our fact verification task in zero-shot scenario. This method uses the same prompt structure as shown in Figure 2 without the closed-domain and the open-domain examples.
- **CICL** (Long et al., 2023; Li et al., 2023): This refers to our closed-domain approach (Equation 2) of the standard ICL workflow that makes use of the labeled data from the FEVER dataset to answer the validity of claims from the other three domains, namely Climate-FEVER, SciFact and Covid-C. With reference to Figure 2, this method uses only the top-white segment in the prompt. The similarity function in this method matches the current input claim $\mathbf{x}$ with only the claims (discarding the evidence part in the matching process) from the training set.
- **CICL-E**: This is similar to CICL with the only difference that both claims and evidences (thus the suffix '-E') are considered to compute the topical similarity used to construct the neighborhood $\mathcal{N}_k(\mathbf{x})$ of Equation 2. Both CICLand CICL-E uses BM25 as the (sparse) similarity computation function.
- **OICL** (Labruna et al., 2024): This baseline refers to the use of unlabeled data from Wikipedia as the additional context used for predicted label generation via Llama-2 (70B). As retrievable units, we use sentences and employ a BM25 based retrieval to obtain the top-m (Equation 3) set of candidate evidences for a query formulated from the input claim $\mathbf{x}$. With reference to Figure 2, this method uses the bottom white segment of the prompt structure (not the top one). Note that there is no '-E' version of this method as for CICL, because the retrieved text is not structured as claim-evidence pairs.

Parametric baselines. To compare our approach with the standard parametric learning approaches, we employ the following baselines:

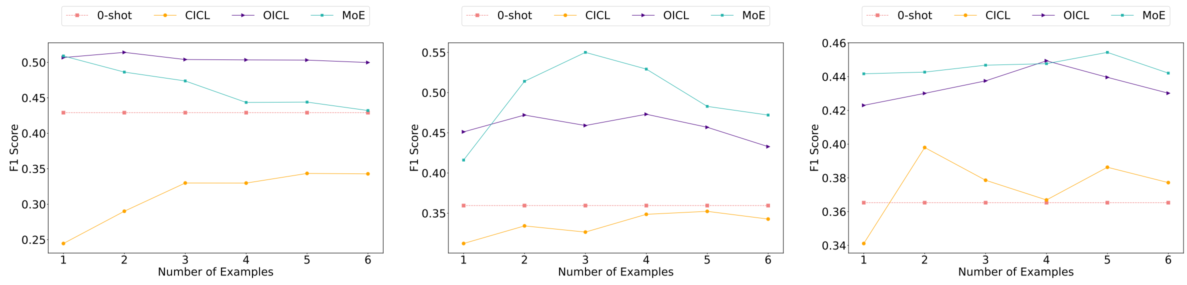
- **RoBERTa** (Long et al., 2023): We finetune RoBERTa (Liu et al., 2019) on the FEVER training data, use this model for prediction on the three target datasets for claim verification.
- **LLaMA-LoRA** (Long et al., 2023; Labruna et al., 2024): We fine-tune the foundation LLM of our non-parametric based approaches, i.e., LLAMA-2 via the low rank domain adaptation technique - LoRA, which in addition to retaining the pretrained weights incorporates additional trainable rank decomposition matrices into each attention layer of a transformer for the purpose of domain adaptation.

Variants of our proposed approaches. We employ two different variants of our proposed MoE-based approach (Equation 4). Similar to CICL, the first variant (which we call **MoE**) uses only claims from the training set to match the current input, whereas the other variant (which we call **MoE-E**) computes the similarity of the current input claim with both claims and their associated evidences from the training set.

5 Results and Analysis

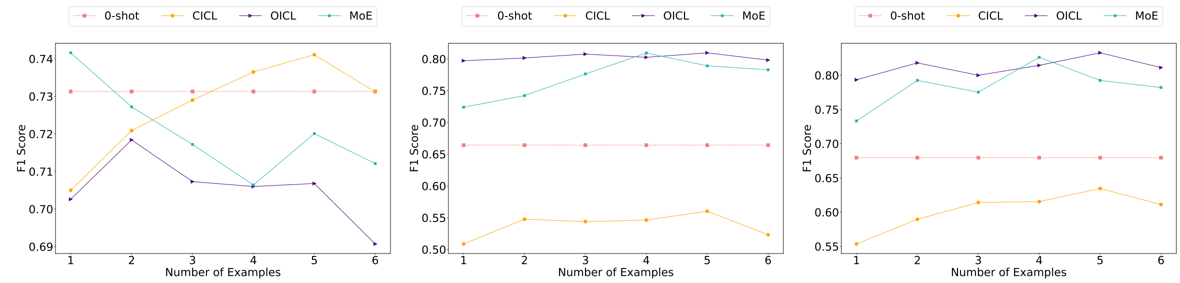
5.1 Main Observations

To compare our approach with the baselines, in Equation 4 we set $\alpha = 0.5$, i.e., we consider equal contributions from both labeled and unlabelled data sources we vary these parameters to see their effect



(a) Climate Fever (b) SciFact (c) Covid C

Figure 3: Sensitivity of the number of examples (labeled or unlabelled) on the various LLM-based approaches for fact verification with the 3-way **SRN setup**.



(a) Climate Fever (b) SciFact (c) Covid C

Figure 4: Sensitivity of the number of examples (labeled or unlabelled) on the various LLM-based approaches for fact verification with the 2-way **SR setup**.

on MoE). To obtain the upper bound on the number of examples (M of Equation 4), we conduct a grid search over a range of $M = 1$ to 10, and report the optimal results for each dataset.

Table 2 presents the results for both the 3-way and the 2-way setups. In general, we observe that non-parametric approaches (even the baseline ones, e.g., OICL, CICL etc.) work more effectively for this task of cross-domain claim verification (which is a novel observation in itself). Generally speaking, using evidences is does not turn out to be effective as can be seen by comparing the results of approaches with and without the suffix '-E'.

We observe that the proposed MoE-based approach mostly outperforms their individual counterparts, i.e., CICL and OICLin terms of overall accuracy and F-score, which indicates that there are useful signals that can be leveraged from both the labeled and the unlabeled data sources. It is likely that the combination method allows one of the approaches to help in prediction when the other one does not turn out to be useful.

It is particularly interesting to see that the non-parametric approaches mostly outperform the supervised ones. Although low rank approximation (Hu et al., 2021) has been reported to work well with few-shot domain transfer (i.e., when a small amount of training data is available for the target domain), in the context of our study it is found to not work well for zero-shot domain transfer (i.e., when no training data is available). Different from parametric approaches, the labeled examples are not tightly integrated to a non-parametric model, which likely allows it to model the desired semantic relationship between claims and evidences in a domain-independent manner.

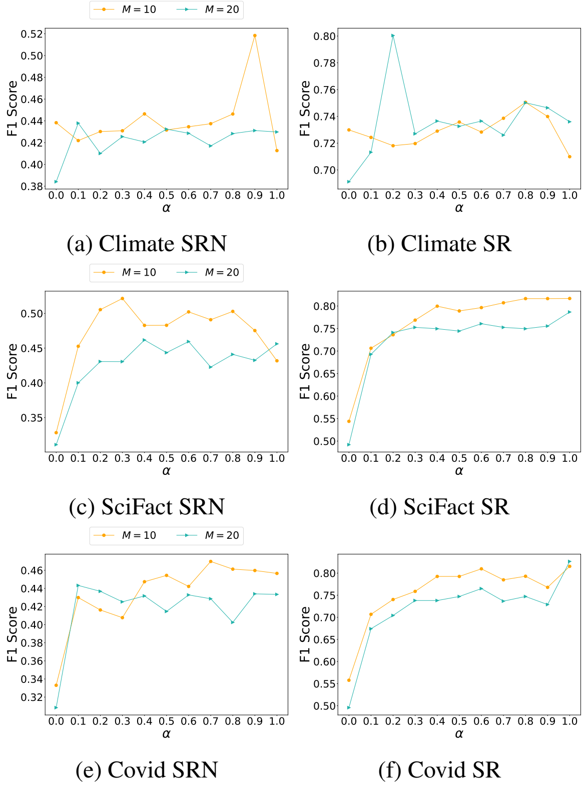
5.2 Sensitivity Analysis

Impact of labeled or unlabelled samples. In this section, we first explore the effect of the number of labeled or unlabelled samples on the non-parametric approaches. For this comparison, for the MoE model we take equal proportion of labeled and unlabelled data as in Table 2. Figure 3 and Figure 4 demonstrate that OICL (unlabelled data as contexts) is relatively more stable and better in performance than CICL (labeled data as example source). Equal contributions of both (as per the MoE approach with $\alpha = 0.5$) turn out to be mostly outperforming the individual methods on SciFact and Covid-C.

Impact of disproportionate contributions from labeled and unlabelled data α on MoE perfor-

mance. Figure 5 reports the effects of varying the relative proportion of labeled vs. unlabelled data on the performance of the mixture model (Equation 4). In general, we observe that the optimal value of α depends on the target domain itself and also largely on the maximum number of candidate examples on which the proportions are defined. For instance, while a high value of α close to 1 turns out to be optimal with $M = 10$ for the SRN-based evaluation on the climate target domain, with $M = 20$ even a lower value of α yields effective results. A likely reason for this is that with higher values of M, the method ends up selecting more labeled data, which due to its out-of-domain characteristics (FEVER vs. Climate) may turn out to be not beneficial for prediction.

Another interesting observation is that the sensitivity analysis shows that downstream performance of fact verification is usually better with disproportionate contributions from labeled and unlabelled data (the optimal points of the plots in Figure 5 either occur to the left or right of the mid-point of the x-axis). This indicates a promising research direction of estimating the desired proportion for specific target domains and even on a per-instance basis.



(a) Climate SRN (b) Climate SR

(c) SciFact SRN (d) SciFact SR

(e) Covid SRN (f) Covid SR

Figure 5: Sensitivity of the proposed mixture model (MoE) on the relative proportion of in-domain (labeled) vs. out-domain (unlabelled) data. As upper bounds of the number of examples, we use $M = 10$ and $M = 20$ (Equation 4).

6 Conclusions and Future work

In this paper, we presented a novel fusion-based approach to combine sources of labeled examples from an out-of-domain training set, and of unlabeled data as additional contexts from a target collection of documents to address the task of zero-shot domain adaptation (no training data available for the target domain) for fact verification. Our method provides a general framework to combine these two sources of data (out-of-domain training vs. Wikipedia) in variable proportions. Our experiments reveal that a carefully tuned proportion of these two different sources of data can provide useful contexts for an LLM-based inference of fact verification.

In the future, we would like to explore ways of predicting the optimal relative proportion of these two sources for our mixture-model for a given target domain. As a part of the adaptation process, we also would like to explore a dynamic choice of this relative proportion based on the current instance (topic of a given claim).

7 Limitations

The major limitation of this work is that we have not used a separate validation set to predict the optimum value of the relative proportion parameter for the proposed MoE model; the reason being, we wanted to investigate a completely zero-shot setup with no availability of a set with ground-truth data. However, we observed that the performance of the MoE-based model is somewhat sensitive to the optimal choice of α (although $\alpha = 0.5$ still outperforms the baselines). A practical application of this model would ideally require a small amount of ground-truth data for tuning this relative proportion for a target domain.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. *arXiv preprint arXiv:2205.12689*.

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10:746–763.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in NIPS*, 33:1877–1901.

Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. 2007. Bias and the limits of pooling for large collections. *Information retrieval*, 10:491–508.

Xinran Chen, Sei-Ching Joanna Sin, Yin-Leng Theng, and Chei Sian Lee. 2015. Why do social media users share misinformation? In *Proceedings of the 15th ACM/IEEE-CS joint conference on digital libraries*, pages 111–114.

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.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Ruoyuan Gao, Yingqiang Ge, and Chirag Shah. 2022. Fair: Fairness-aware information retrieval evaluation. *Journal of the Association for Information Science and Technology*, 73(10):1461–1473.

Aryo Pradipta Gema, Pasquale Minervini, Luke Daines, Tom Hope, and Beatrice Alex. 2024. Parameter-efficient fine-tuning of llama for the clinical domain. *Preprint*, arXiv:2307.03042.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. *arXiv preprint arXiv:2104.07467*.

Christopher Hidey and Mona Diab. 2018. Team sweeper: Joint sentence extraction and fact checking with pointer networks. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 150–155.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring listwise evidence reasoning with t5 for fact verification. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP (Volume 2: Short Papers)*, pages 402–410.

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. *arXiv preprint arXiv:2011.03088*.

Jihyuk Kim, Minsoo Kim, Joonsuk Park, and Seung-won Hwang. 2023. Relevance-assisted generation for robust zero-shot retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 723–731, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Tiziano Labruna, Jon Ander Campos, and Gorka Azkune. 2024. When to retrieve: Teaching llms to utilize information retrieval effectively. *arXiv preprint arXiv:2404.19705*.

Nineli Lashkarashvili, Wen Wu, Guangzhi Sun, and Philip C. Woodland. 2024. Parameter efficient fine-tuning for speech emotion recognition and domain adaptation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in NIPS*, 33:9459–9474.

Xiaoqian Li, Ercong Nie, and Sheng Liang. 2023. From classification to generation: Insights into crosslingual retrieval augmented icl. *arXiv preprint arXiv:2311.06595*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

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*.

Quanyu Long, Tianze Luo, Wenya Wang, and Sinno Jialin Pan. 2022. Domain confused contrastive learning for unsupervised domain adaptation. *arXiv preprint arXiv:2207.04564*.

Quanyu Long, Wenya Wang, and Sinno Jialin Pan. 2023. Adapt in contexts: Retrieval-augmented domain adaptation via in-context learning. *arXiv preprint arXiv:2311.11551*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Liangming Pan, Yunxiang Zhang, and Min-Yen Kan. 2023. Investigating zero-and few-shot generalization in fact verification. *arXiv preprint arXiv:2309.09444*.

Andrew Parry, Sean MacAvaney, and Debasis Ganguly. 2024. Exploiting positional bias for query-agnostic generative content in search. *CoRR*, abs/2405.00469.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. *arXiv preprint arXiv:2106.03794*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *arXiv preprint arXiv:2103.08541*.

Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*.

Dominik Stammbach. 2021. Evidence selection as a token-level prediction task. In *Proceedings of the Fourth Workshop on FEVER*, pages 14–20, Dominican Republic. ACL.

Dominik Stammbach and Guenter Neumann. 2019a. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109, Hong Kong, China. Association for Computational Linguistics.

Dominik Stammbach and Guenter Neumann. 2019b. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on FEVER*, pages 105–109, Hong Kong, China. ACL.

Shyam Subramanian and Kyumin Lee. 2020. Hierarchical Evidence Set Modeling for automated fact extraction and verification. In *Proceedings of the 2020 Conference EMNLP*, pages 7798–7809, Online. ACL.

Alon Talmor and Jonathan Berant. 2019. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. *arXiv preprint arXiv:1905.13453*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2022. Gpt-j-6b: A 6 billion parameter autoregressive language model, 2021.

Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. Check-covid: Fact-checking covid-19 news claims with scientific evidence. *arXiv preprint arXiv:2305.18265*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in NIPS*, 35:24824–24837.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148*.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Transformer-xh: Multi-evidence reasoning with extra hop attention.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. Reasoning over semantic-level graph for fact checking. *arXiv preprint arXiv:1909.03745*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043.

	Random Neighbour			BM25 Neighbour		
	CICL		OICL	CICL		OICL
Setting	w-Evi	w/o-Evi	w/o-Evi	w-Evi	w/o-Evi	w/o-Evi
SRN	**0.4127**	0.4486	0.4292	0.3434	**0.4559**	**0.4503**
SR	0.4619	0.4379	0.4619	**0.4685**	**0.7411**	**0.7184**

Table 3: The impact of randomly selected examples on Climate Fever dataset, F1 score

A Appendix

A.1 Comparative Analysis: RoBERTa vs. MoE

Table 2 in shows that in SRN setup, our proposed methodology MoE consistently outperforms RoBERTa in terms of F1 score for all the three datasets. More specifically MoE shows a substantial improvement of 55.01% F1 score indicating MoE's enhanced ability to adapt to climate-related claims through dynamic context selection. To further substantiate the advantages of our proposed method over RoBERTa, we present Table 4 which showcases specific claims where MoE successfully validates the information, whereas RoBERTa fails. A possible intuition is that RoBERTa shows limited generalizability for our downstream task due to its dependence on domain-specific training data.

A.2 Sensitivity Analysis

More detailed results of the sensitivity of MoE performance to different α values, representing the balance between CICL and OICL contexts, in SRN and SR setups on the SciFact and Covid C datasets in terms of F1 scores and accuracy are depicted in Table 5.

A.3 Performance of ICL-based frameworks with Randomly Selected Demonstrations

To concretely conclude, we additionally investigate whether the randomly selected demonstrations can

Index	Climate Fever - Claim	GT label	MoE	RoBERTa
1	Global warming is driving polar bears toward extinction.	Supports	Supports	NEI
2	Ice berg melts, ocean level remains the same.	Refutes	Refutes	NEI
3	Sea level rise is not going to happen.	Refutes	Refutes	NEI
4	CO2 changes are closely related to temperature.	Supports	Supports	NEI
5	The ovary is an organ involved in the creation of new life.	Supports	Supports	NEI
6	The contribution of waste heat to the global climate is 0.028 W/m2.	Supports	Supports	NEI
7	Venus is not hot because of a runaway greenhouse.	Refutes	Refutes	Supports

Table 4: Examples showcasing our proposed methodology MoE's superior performance over supervised baseline RoBERTa for Climate Fever dataset in OOD fact verification task.

		SRN settings						SR settings					
		Climate FEVER		SciFact		COVID C		Climate FEVER		SciFact		COVID C	
α	Metric ↓	M → 10	20	10	20	10	20	10	20	10	20	10	20
0.0	**Acc**	0.5220	0.5033	0.4433	0.4367	0.3833	0.0.3722	**0.7949**	0.7839	0.6862	0.6755	0.5868	0.5454
	F1	0.4383	0.3941	0.3283	0.3110	0.3330	0.3083	0.7299	0.6912	0.5439	0.4920	0.5577	0.4957
0.1	**Acc**	0.5206	0.5308	0.5167	0.4967	0.4667	0.4889	0.7883	0.7905	0.7766	0.7713	0.7107	0.6859
	F1	0.4219	0.4379	0.4527	0.4001	0.4299	**0.4433**	0.7243	0.7132	0.7063	0.6928	0.7068	0.6742
0.2	**Acc**	**0.5314**	0.5185	0.5567	0.5102	0.4556	0.4778	0.7861	0.7246	0.7926	0.7979	0.7438	0.7107
	F1	0.4302	0.4101	0.5054	0.4307	0.4162	0.4368	0.7181	<u>**0.8004**</u>	0.7360	0.7412	0.7403	0.7042
0.3	**Acc**	0.5170	0.5278	<u>**0.5633**</u>	0.5100	0.4556	0.4778	0.7795	0.8037	0.8085	0.8032	0.7603	0.7438
	F1	0.4398	0.4255	<u>**0.5214**</u>	0.4306	0.4077	0.4252	0.7197	0.7270	0.7687	0.7526	0.7587	0.7380
0.4	**Acc**	0.5272	0.5257	0.5333	**0.5310**	0.4833	0.4778	0.7828	0.8049	0.8351	0.8032	0.7934	0.7438
	F1	0.4462	0.4206	0.4829	**0.4617**	0.4474	0.4317	0.7290	0.7365	0.7997	0.7496	0.7925	0.7380
0.5	**Acc**	0.5177	<u>**0.5358**</u>	0.5333	0.5233	0.5011	0.4667	0.7872	0.8060	0.8245	0.7979	0.7924	0.7521
	F1	0.4316	**0.4324**	0.4829	0.4435	0.4544	0.4146	0.7358	0.7326	0.7891	0.7444	0.7925	0.7471
0.6	**Acc**	0.5156	0.5344	0.5467	0.5302	0.4833	0.4833	0.7806	0.8049	0.8298	0.8085	0.8099	0.7686
	F1	0.4346	0.4286	0.5022	0.4595	0.4422	0.4328	0.7283	0.7365	0.7965	0.7607	0.8097	0.7650
0.7	**Acc**	0.5127	0.5243	0.5401	0.5101	0.5056	0.4833	0.7850	0.7949	0.8404	0.8032	0.7851	0.7438
	F1	0.4374	0.4170	0.4909	0.4227	<u>**0.4699**</u>	0.4286	0.7386	0.7260	0.8072	0.7526	0.7848	0.7366
0.8	**Acc**	0.5177	0.5344	0.5500	0.5200	0.5111	0.4778	0.7927	<u>**0.8082**</u>	<u>**0.8457**</u>	0.8032	0.7924	0.7421
	F1	<u>**0.4464**</u>	0.4283	0.5029	0.4410	0.4614	0.4025	**0.7505**	0.7502	<u>**0.8165**</u>	0.7496	0.7929	0.7471
0.9	**Acc**	0.5185	0.5315	0.5300	0.5133	<u>**0.5167**</u>	**0.4944**	0.7828	0.7982	<u>**0.8457**</u>	0.8032	0.7686	0.7355
	F1	0.4463	0.4311	0.4753	0.4325	0.4598	0.4340	0.7399	0.7464	0.8165	0.7555	0.7678	0.7289
1.0	**Acc**	0.4989	0.5040	0.5303	0.5267	0.5722	0.4589	0.7354	0.7663	0.8297	**0.8138**	**0.8182**	<u>**0.8264**</u>
	F1	0.4127	0.4298	0.4318	0.4562	0.4566	0.4334	0.7099	0.7360	0.8168	**0.7867**	**0.8153**	<u>**0.8264**</u>

Table 5: Detailed results of MoEperformance varying the value of α across all the 3 datasets in SRN and SR Setups. The best results are shown in bold and the dataset-specific best results in each setup SR and SRN have been underlined.

enhance the prediction performances of CICL and OICL based frameworks. From Table 5 we can observe that **CICL** and **OICL with randomly selected examples** setups perform worse compared to their specific counterparts (CICL and OICL). This is evident in both SRN and SR settings where the F1 scores are lower for randomly selected examples. The likely reason behind this performance is that randomly selected examples may not provide the most relevant context or may include irrelevant information. This lack of targeted context likely contributes to the lower F1 scores. The model benefits more from carefully selected examples that are specifically relevant to the claims being verified. Additionally, from Table 5 we can get another finding that evidence-aware setup tends to result in better performance than evidence-agnostic setup, but the improvement is not substantial for randomly selected setups. The general perception behind it is that an evidence-aware setup generally helps the model by providing additional context that is directly relevant to the claim. This context aids in better understanding and verification of the claims, leading to improved performance in random CICL set up. Hence, the presence of evidence helps in grounding the model's predictions more firmly in random CICL and OICL setups.

A.4 Information of our Annotated data

Table 6 explains that during annotations, experts labeled the claims as NEI, however, we have annotated those NEI claims as support/refute, which proves the existence of noisy data in their annotations.

FEVER - Claim	Evidence	GT label	Annotated
The television series Fringe is available on DVD.	"Fringe," the television series, was available on DVD. However, availability can vary by region and over time, so it's recommended to check with local retailers or online platforms to confirm current availability in your area.	NEI	Supports
Soul Food is the only comedy film to ever exist.	Certainly not! While "Soul Food" is a popular comedy-drama film, it's just one among thousands of comedy films that have been produced over the years.	NEI	Refutes
Mary of Teck's son abdicated the throne in 1902.	Opposition to her third husband Bothwell led to the formation of a coalition of nobles, who captured Mary and forced her to abdicate in favor of her son, who came to the throne as James VI in 1567.	NEI	Refutes
Ragtime features Samuel L. Jackson as a dancer.	In the movie "Ragtime" released in 1981, Samuel L. Jackson did appear in a role, but he was not specifically cast as a dancer in that film.	NEI	Refutes
Black Canary is a character in Batman comic books.	Black Canary has been adapted into various media, including direct-to-video animated films, video games, and both live-action and animated television series, featuring as a main or recurring character in the shows Birds of Prey, Justice League Unlimited, Smallville, Batman: The Brave and the Bold, Young Justice and Arrow.	NEI	Supports
The ovary is an organ involved in the creation of new life.	It is an inflammatory mass involving the fallopian tube, ovary and, occasionally, other adjacent pelvic organs.	NEI	Supports
There are stripes on the Bengal tiger.	Such a tiger has the black stripes typical of the Bengal tiger, but carries a white or near-white coat.	NEI	Supports
Saturn Corporation is a subsidiary of Disney.	The Saturn Corporation, also known as Saturn LLC, is a registered trademark established on January 7, 1985, as a subsidiary of General Motors	NEI	Refutes

Table 6: Comparison between the ground truth (GT) annotations and our actual annotations (Annotated) for claims labelled as 'Unverifiable,' extracted from the FEVER dataset.