PEINet: Joint Prompt and Evidence Inference Network via Language Family Policy for Zero-Shot Multilingual Fact Checking

Xiaoyu Li 1,2,*,†, Weihong Wang 2,‡, Jifei Fang 1,2, Li Jin 1,2,*‡, Hankun Kang 1,2 and Chunbo Liu 1,2

1 Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
2 Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
* Correspondence: lixy01@aircas.ac.cn (X.L.); jinlimails@gmail.com (L.J.)
† These authors contributed equally to this work.
‡ Work done as an intern at NIST.

Abstract: Zero-shot multilingual fact-checking, which aims to discover and infer subtle clues from the retrieved relevant evidence to verify the given claim in cross-language and cross-domain scenarios, is crucial for optimizing a free, trusted, wholesome global network environment. Previous works have made enlightening and practical explorations in claim verification, while the zero-shot multilingual task faces new challenging gap issues: neglecting authenticity-dependent learning between multilingual claims, lacking heuristic checking, and a bottleneck of insufficient evidence. To alleviate these gaps, a novel Joint Prompt and Evidence Inference Network (PEINet) is proposed to verify the multilingual claim according to the human fact-checking cognitive paradigm. In detail, firstly, we leverage the language family encoding mechanism to strengthen knowledge transfer among multi-language claims. Then, the prompt turning module is designed to infer the falsity of the fact, and further, sufficient fine-grained evidence is extracted and aggregated based on a recursive graph attention network to verify the claim again. Finally, we build a unified inference framework via multi-task learning for final fact verification. The newly achieved state-of-the-art performance on the released challenging benchmark dataset that includes not only an out-of-domain test, but also a zero-shot test, proves the effectiveness of our framework, and further analysis demonstrates the superiority of our PEINet in multilingual claim verification and inference, especially in the zero-shot scenario.

Keywords: multilingual fact-checking; zero-shot; multi-language bert; language family; prompt learning; recursive graph attention network; multi-task learning

1. Introduction

Misinformation is low-cost, wide-spread, and large-scale replicable on social media, which has significantly influenced democracy, justice, and public trust [1]. Especially with the support of new multimedia technology, such as short videos on social media, misinformation spread has transcended countries and languages. Research indicates that the proportion of false claims is only approximately 1% of total news dissemination on all platforms, while the percentage of false claims is almost more than 6% of total tweets on social media [2]. Human fact-checking is of high-quality but is time-consuming, it refers to specifically confronting false information with a deceptive purpose, such as offensive news, political deception, and dramatic online rumors around the world. The absence of professional auditors, particularly who understand minority languages, further increases the challenge of multilingual fact-checking. Thus, multilingual automated fact-checking has become urgent to research, including the release of multilingual fact-checking datasets and inventing computational approaches.
Multilingual fact-checking is a special fact-checking task that aims to leverage related evidence from a multilingual text corpus to verify a multilingual textual claim. Figure 1 for example, shows two such samples from Portuguese and Spanish. In the special task, how to effectively retrieve, align, integrate, infer, and verify is the major challenge and still remains an open question, especially leveraging resources from different languages in multilingual scenes.

| Claim | Evidence | Label |
|-------|----------|-------|
| #1 O lunático do Dória tomando a vacina chinesa, enfermeira sem luva e os dois sem máscara. Imagina se fosse o Bolsonaro? | #E1 #1 The photo was taken on the first day of the national flu vaccination campaign, promoted by the Ministry of Health. | Mostly-False, Mostly-False, Complicated/Hard to Categorise, Other |
| #2 Bossio habló un incremento en la matrícula secundaria del 20%.” | #E2 #2 This Tuesday (21), 9000 Brazilian health professionals, in several states, will receive doses of the vaccine and will be monitored to assess whether the substance is efficient and safe. | Mostly-False, Mostly-False, Complicated/Hard to Categorise, Other |
| #3 Lupa was already circulating before the Start of the tests, but it was shared again this week. It's not the only piece of disinformation published this week. #E3 | | Mostly-False, Mostly-False, Complicated/Hard to Categorise, Other |

Figure 1. Typical Examples from X-FACT. Example #1 and #2 are from Portuguese and Spanish, respectively. Every sample consists of a claim, evidence, and label. Due to the lack of space, the original claims are only shown in the figure, and the translations (symbol #1) of claims and parts of evidence are offered as the central part to better illustrate the evidence.

In previous studies, a possible approach for multilingual verification was to utilize translation systems (e.g., Google Translator) for news verification based on multilingual evidence, which hypothesized that fake news tended to be less widespread across languages as compared to authoritative news [3]. However, the threshold value of the modeling pipeline needed to be set manually. In consequence, the end-to-end frameworks based on multilingual fine-tuning transformers were designed to encode multilingual claims and evidence for fact-checking. Schwarz, S et al. [4] proposed the EMET framework based on a convolutional neural network to classify the reliability of messages posted on social media platforms. Shahi and Nandini [5] built a BERT classifier to detect the false class and other fact-check articles. Roy, A. and Ekbal, A. [6] next proposed an automated multi-modal content approach based on the multilingual fact verification system, which automated the task of fact-verifying websites, and provided evidence for each judgment. Nevertheless, they merely utilized simple evidence combination methods such as concatenating the evidence or just dealing with each evidence-claim pair.

Going further, in addition to considering the use of pre-trained models to achieve multilingual semantics, another promising direction is to distill knowledge from high-resource to low-resource languages. Kazemi, A. et al. [3] adopted the teacher–student framework to train a multilingual embedding model based on the XLM-R model, in which a high-quality teacher model is adopted to promote the learning ability of a student model. With the expansion of language categories, a potentially troublesome problem in fine-grained multilingual fact-checking in this way is that the teacher model has maybe never seen the statements expressed in a minority language. Thus, the intensive research of few-shot or zero-shot multilingual fact-checking has been subject to considerable discussion. Martín, A. et al. [7] proposed FaceTeR-Check architecture for semi-automated fact-checking, which enables several modules to evaluate semantic similarity to calculate natural language inference and to retrieve information from online social networks. The multilingual fact-
checking tool could verify new claims, extract related evidence, and track the evolution. Lee, N. et al. [8] leveraged an evidence-conditioned perplexity score from the masked language model for claim verification, which is an unsupervised method. In the fine-gained scenario, Gupta, A. and Srikumar, V. [9] determined the veracity of a claim by adopting an attention-based evidence aggregator that used a multilingual transformer-based model, and concatenated the additional metadata of a claim to augment the model.

The unified models based on pre-training and fine-tuning are the mainstream fact-checking paradigm, for example, designing the evidence-conditioned perplexity function or evidence aggregator based on the attention mechanism for fact verification, which is mainly trained for learning representation patterns and joint probability distributions of factual claims and pieces of evidence. The methods indicated a feasible inference idea; the core of this is that with deep semantic understanding from multilingual pre-training models, a downstream classifier can be progressively fine-tuned based on a training dataset in a certain scale. In consequence, sufficient evidence of aggregation and metadata supplementation for fact verification contributes to strengthening the authenticity inference calculation. However, in the scenario of few-shot or zero-shot samples across domains and languages, the learning perception ability of the previous models is extremely limited. The main reasons are as follows: On the one hand, due to the lack of large-scale data driving, the model cannot be quickly and finely fine-tuned to the downstream tasks, and the potential of pre-training models is desired to be further stimulated. On the other hand, the existing coarse-grained evidence aggregation mechanism, which requires more fine-grained evidence chains, cannot provide acceptable evidence. Last, but not the least, the previous models have not considered leveraging the consistency of the same language family in the fact statement representation paradigm to deal with the correlation learning mechanism of the lifting model, thus, strengthening the transfer learning ability of the model in few-/zero-shot scenarios.

To deal with the above issues, we propose the Joint Prompt and Evidence Inference Network (henceforth, PEINet) framework for multilingual fact-checking in the zero-shot scene, which is inspired by the human cognitive paradigm. When the human brain judges whether a fact is true or false, the first decision-making consciousness is to take aim at the claim or title. If there is an excessively exaggerated or obviously outrageous expression in the statement, we can directly determine that it is false. Of course, for an obscure and complicated statement, next it is necessary to retrieve relevant evidence and extract reasonable clues, so as to generate final judgment. Furthermore, in the face of fact-checking in terra incognita, relevance cognition plays an important role in the brain learning mechanism. In general, claims in a similar-expression paradigm may have consistent authenticity, and similar fact statements described in the same language family may share the like expression paradigms.

In order to better clarify the mechanism, we take a concrete example, as shown in Figure 1. Through comprehensive comparison and observation of examples #1 and #2, we discover that the claim in #1 obviously contains exaggerated and deliberate contents that are highlighted with roughening and underlining in the figure, which evidently is false from logical empiricism. Furthermore, the claim in #2 is relatively objective. Next, as shown in the purple dotted lines and order numbers in circles in the figure, the fine-grained clue chains of evidence in red font explored from the claim and retrieved evidence can verify that the claim is clearly false. Further consideration is given to the fact that Portuguese and Spanish belong to the common language family ‘Indo-Aryan: Romance’. By strengthening the interaction of different claims in the same language family, the model can transfer knowledge between the claims. In other words, even if the model only is trained with the Portuguese claims, the Spanish claims in the test can be verified by the model. Specifically, in PEINet, we first design a language family-shared embedding layer to strengthen the model for the inter-semantic understanding between different claims belonging to the same language family. Then, we design a co-interactive cognitive inference layer, in which a prompt-based claim verification module is responsible for directly judging the claim,
and the fine-grained evidence aggregating module takes the chance of inferring the deep-seated evidence clues. Finally, to achieve a unified cognitive architecture, the multi-task learning mechanism is adopted for building the affine classification layer. The PEINet model, as a novel algorithm, can be adapted to the multilingual fact-checking social media platform to automatically identify and filter misinformation, fake news, political deception, online rumors, etc. For the information published by the social media platforms, the system automatically submits it to the algorithm for review. Then, the algorithm can verify the claim and choose whether to block it or strictly review it manually according to the classification label. In particular, when a piece of misinformation is refuted, and is deliberately repackaged in another language, the algorithm can replace manual audits efficiently and greatly relieve the pressure on auditors.

Experiments on the challenging real-world datasets X-Fact demonstrate that our model outperforms several state-of-the-art scores in in-domain, out-of-domain, and zero-shot tests in terms of the macro-F1 evaluation metric. Moreover, the comprehensive experimental analysis confirms the validity and the reasonableness of our model. To sum up, our main contributions are as follows:

- We propose PEINet, a novel framework inspired by the human cognitive decision paradigm to verify multilingual claims in out-of-domain and zero-shot scenarios.
- We explore that the shared encoding mechanism via language family metadata augmentation strengthens authenticity-dependent learning between cross-language claims. Moreover, we investigate the multilingual fact-checking model that integrates prompt-based judgment and further fine-grained evidence aggregation inferences for the final claim verification based on the multi-task learning mechanism.
- We perform exhaustive empirical evaluation and ablation studies to demonstrate our model’s effectiveness, especially in zero-shot scenarios, and further discuss the potential optimizable issues.

2. Related Work

There are extensive studies about misinformation detection from multiple perspectives; these studies derive several analogous problems such as rumor detection and verification, fake news detection, machine-generated news, false information detection, and fact-checking [10]. Due to this important research, exciting advances have been achieved in automated fact-checking with increasingly larger datasets, better performing models, and more powerful systems. However, automated fact-checking is still a challenging task unquestionably, especially in a multilingual or multi-modal scenario. Engineering accurate systems and advanced computational methods for fact-checking are required urgently [11].

2.1. Automated Fact-Checking

A dominant bottleneck associated with fact-checking is the cause of the shortage of available research datasets. With the rise of fact-checking websites [12], corpus annotation platforms, and pre-trained language models [13], there has been an increase in the volume of available data [10]. The LIAR dataset, which consists of 12.8K claims retrieved from Politifact API, was first released for training machine learning models. The meta-data (i.e., speaker of claim, political affiliation, medium by the claim that was first issued) for the claims was proactively annotated. However, the evidence applied by fact-checkers for verifying the claims was not delivered in the dataset [14]. Urged by the announcement of LIAR, huge strides in progress have been made, and there have been substantially larger datasets, in which the FEVER [15], FEVER 2.0 [16] and FEVEROUS [17] datasets have been the landmark tasks. An essential justification method of claims was to reveal which evidence was judged to reach a verdict, as was required in FEVER-series tasks. The normal framework for auto-fact-checking comprised of three sub-tasks: (i) worthy claim detection to identify the claim which is needed to be verified; (ii) retrieving documents related to the claim to select the most relevant pieces of evidence; (iii) predicting the veracity of the claim with explanation extraction [10,11,18]. A combination of heuristic
approaches, machine learning methods, and deep neural network models are adopted as a pipeline model (e.g., a document retrieval module based on TF-IDF [19], neural semantic matching networks for combining fact extraction [20], and evidence inference networks for interpretable claim verification [21]) to solve the sub-task of each stage for the implied task of label aggregation or the joint model (e.g., end-to-end system to assess the veracity of claims [22]) to verify the claim.

2.2. Multilingual Cross-Domain Scenario

Although different theories, models, and systems in research among misinformation have transcended countries and languages [23–25], most investigations have focused on English [26,27] or other monolingual datasets [28]. There is yet a pressing challenge in multilingual fact-checking—relevant datasets released for performance evaluation of multilingual fact-checking models in more languages have been promoted for breaking through the bottleneck [29,30], and have accelerated progress without the dependence on translation systems [31]. For instance, the multilingual cross-domain dataset of 5182 fact-checked news articles for COVID-19 was introduced by Shahi and Nandini [5]. The larger publicly available multilingual dataset X-FACT for factual verification of naturally existing real-world claims was issued afterwards, which contained 31,189 general domain non-English claims from 25 languages [9]. Recently, a large-scale multilingual and multi-modal fact-checked dataset collected from social network contains rich social media data in 41 different languages, and was firstly made available as a heterogeneous graph [30].

Regarding the above examples, this provides further motivation that auto-fact-checking prospective models are desired for applying to multilingual cross-domain scenarios (e.g., public health, society, or politics). How to effectively achieve semantic comprehension and leverage resources between different languages is still an open question, especially when handling low-resource languages. To the extent that is known, the few-shot learning or zero-shot learning method has been successfully applied for the low-resource task, which discovered that pre-trained language models (LMs) have stored factual knowledge in their parameters, and has advantages in the comprehension of samples not included in the training dataset [32]. Therefore, leveraging the few-/zero-shot ability of multilingual LMs for fact verification is a feasible strategy in multilingual scenarios.

2.3. Zero-Shot Inference Verification

The large-scale auto-regressive LMs were adjusted through few-shot and zero-shot learning for wide-ranging tasks with obviously less computational cost than full fine-tuning [33]. Regarding most monolingual pre-trained models, multilingual-masked and sequence-to-sequence models that perform cross-lingual tasks with fine-tuning on the labeled data in downstream tasks have been designed, such as mBERT, XLM-R, and mT5 [34]. Considering that multilingual few-shot learning capabilities are less well understood in the above models, the comprehensive multilingual models (as GPT-3 [35] and XGLM [34]) were proposed in the few-shot and zero-shot learning paradigm for multilingual natural language understanding tasks without fine-tuning. However, the multilingual models can only generate a fundamental semantic understanding of claims and evidence, and the abilities of multilingual fact-checking mainly depend on downstream module designs.

For further leveraging the powerful zero-shot learning abilities of LMs, a new way that takes advantage of LMs (including BERT, GPT2) via a perplexity score was shown to exhibit zero-shot capabilities, thus, outperforming the Major Class baseline models by more than 10% on F1-Macro metrics across multiple datasets [8]. In addition, two new cross-domain fact-checking datasets corresponding to the model were released to prove the rationality of the perplexity score synchronously. However, the best optimal perplexity threshold that separates supported or unsupported claims was generated from the statistical calculation of evidence-conditioned perplexity scores for each claim, which was not automatically evaluated in models. As an automatic framework, Question Answering for Claim Generation (QACG) [36] was proposed to train a fact-verification model by using automatically gener-
ated claims that can be supported, refuted, or unverified from the evidence. The model improved Roberta’s performance by up to 22% when experimented on the FEVER dataset, and reduced the human-annotated demand that was equivalent in performance to 2K+ manually curated examples in the zero-shot scenario. When solving the problem of the multilingual zero-shot fact-checking tasks, the machine learning classifiers automatically identified COVID-19 misinformation via learning contextualized embedding. The misinformation from social media included English, Bulgarian, and Arabic [37]. As shown in the experimental results, the BERT and multilingual BERT achieved the best result.

Nonetheless, the task only explored contextualized embedding from the LMs for misinformation detection without delving into the cross-lingual implicit value, especially the mutual information transfer between the different languages. In another study, the zero-shot capabilities of several multilingual transformer-based models were evaluated based on the X-Fact dataset for the fact-checking task. Through modeling the attention relationship between the textual claim and evidence from news stories retrieved using a search engine, the attention-based evidence aggregator (Attn-EA) model outperformed the claim-only model (Claim-Only) [9]. However, as a baseline for suggesting the X-Fact dataset, the model—without making full use of additional metadata—lacked the interpretable claim verification and fine-grained evidence aggregating and inferring, and was not complete to construct a unified fact-checking cognitive framework.

3. Methods

In this section, we introduce our PEINet framework for zero-shot multilingual fact-checking, which is the main contribution of this paper. Figure 2 shows the framework, which mainly introduces three components: (1) An LF-aware shared encoding layer (details in Section 3.2), which enhances authenticity-dependent learning between multilingual samples by encoding the language family metadata. (2) A co-interactive cognitive inference layer (details in Section 3.3), in which the prompt-based claim verification module verifies whether the claim is true or false depending on the claim contents. Furthermore, the fine-grained evidence aggregating module obtains recursively valuable evidence to verify the claim again. (3) A unified affine classification layer (details in Section 3.4), which leverages the multi-task learning mechanism to achieve the final classification uniformly.
3.1. Problem Formulation

First, we describe the mathematical modeling of the multilingual fact-checking problem. For a dataset \( T = \{X, Y\} \), in which \( X = \{x_1, x_2, \ldots, x_n\} \) and \( Y = \{y_1, y_2, \ldots, y_m\} \), there are \( n \) samples and \( m \) labels. Giving, arbitrarily, an input \( x_i \in X \), which includes a claim \( c_i \), the corresponding metadata \( m_i \), and \( N \) pieces of retrieved evidence \( E_i = \{e_1, e_2, \ldots, e_N\} \), the model output is the truthfulness label of the claim \( \hat{y}_i \in Y \). We aim to train the model that can fit a more accurate function \( F \) so that the predict label \( \hat{y}_i \) is more probable to be consistent with the true label \( y_i \).

3.2. LF-Aware Shared Encoding Layer

We employ mBERT [38] as the shared encoder by extracting the [CLS] embedding token as the representation, where the [CLS] is a special classification token and the last hidden state of mBERT corresponding to this token (h[CLS]).

Specifically, the claim and retrieved evidences are, respectively, fed into mBERT for obtaining the corresponding semantic embedding vectors. Notably, the feature of language family (LF) is extracted beforehand from the metadata for augmenting the cross-language comprehension ability. The form of the LF feature is a string template modified by special characters [unused*], e.g., \( m = \text{"[unused1], ar, [unused1], [unused2], Afro-Asiatic, [unused2]."} \), where the [unused1] and [unused2] are also the special tokens for fixed mark. Considering there are abnormal characters in the claims and evidence, we also replace these invalid with a null character. Subsequently, the new input is contacted as follows:

\[
\hat{c} = [m, c], \\
\hat{e}_i = [m, e_i].
\] (1)

Next, the input is re-encoded as follows:

\[
x_i = \{\hat{c}_i, \hat{E}_i\} \\
= \{\hat{c}_i, \hat{e}_1, \hat{e}_2, \ldots, \hat{e}_N\}.
\] (2)

3.3. Co-Interactive Cognitive Inference Layer

3.3.1. Prompt-Based Claim Verification Module

As we mentioned before, the prompt is well developed for few-/zero-shot classification tasks such as multi-class fact verification and other natural language inferences [39]. Considering that most auto-generating prompts cannot achieve a comparable performance, we design task-specific manual prompts for fact-checking [40,41]. In this section, we introduce the details of the prompt-based claim verification module, and illustrate how to adapt the prompt-tuning for fact-checking inferences.

As a cloze-style task for tuning PLMs, there are template \( T(\cdot) \) and a set of label words \( V \) formally. For each input instance \( x \), the template is designed to map \( x \) to the prompt input \( x_{\text{prompt}} = T(x) \). Notably, the template defines whether the position of input tokens is needed to be adjusted and whether to pad other additional tokens.

As the task contains seven classification labels, as described in detail in Section 4.1, at least two parameters are required to be placed into the template for filling the label words. Specifically, the template is designed as:

\[
T(X) = \text{The true label of the claim is [MASK1] [MASK2]. } X,
\] (3)

where the \( X \) is the content strings.
Accordingly, the template maps the input of claim sequence $\hat{c}_i$ to $x_{\text{prompt},i}$ which can be formalized as Equation (4):

$$x_{\text{prompt},i} = \{[\text{CLS}], \text{ the true label of the claim is } \text{[MASK1][MASK2]}, +$$

$$\text{language family tokens } [\text{SEP}], w_1, w_2, \cdots, w_{|x_1|}, \cdots, +$$

$$\text{claim tokens } w_1, w_2, \cdots, w_{|x_2|}, \cdots, +[\text{SEP}].\}$$

where the symbol $+$ only represents string concatenation.

Then, we use the mBERT to encode all tokens of the input sequence into corresponding embedding vectors $H = \{h_{[\text{cls}]}, \cdots, h_{[\text{MASK1]}}, h_{[\text{MASK2]}}, h_{[\text{SEP}]}\}$. Next, by inputting $x_{\text{prompt}}$ into mBERT, the embedding vectors $h_{[\text{MASK}]} \in H$ in hidden layers are generated based on the feed-forward calculation. Given $v \in V$, we calculate the probability that the token $v$ is the cloze answer of the masked position, as shown in Equation (5):

$$p([\text{MASK}] = v | x_{\text{prompt}}) = \frac{\exp(v \cdot h_{[\text{MASK}])}}{\sum_{\tilde{v} \in V} \exp(\tilde{v} \cdot h_{[\text{MASK}])}},$$

where $v$ is the embedding vector of the token $v$ in the mBERT.

Further, we bridge the mapping relationship between the set of labels and the set of words, which is expressed as the affine function in Equation (6):

$$\Phi : \mathcal{Y} \rightarrow \mathcal{V}.$$ (6)

As mentioned above, the prompt template contains multiple [MASK] tokens, and the label probability is, thus, considered with all masked positions to make final predictions. The formula is Equation (7):

$$p(y_{\text{prompt}} | x_{\text{prompt}}) = \prod_{j=1}^{n} p([\text{MASK}]_j = \phi_j(y) | T(x_{\text{prompt}})), \quad (7)$$

where $n$ is the number of masked positions in $T(\cdot)$, and $\phi_j(y)$ is to map the class $y$ to the set of label words $\mathcal{V}_{[\text{MASK}]}$ for the $j$-th masked position $[\text{MASK}]_j$. The $y_{\text{prompt}}$ corresponding to the maximum probability value is the prediction label.

Furthermore, the optimization objective in the prompt-based module is to minimize the loss function in Equation (8):

$$\mathcal{L}_{\text{prompt}}(y_{\text{prompt}}, y) = \frac{1}{|X|} \sum_{x \in X} \log \prod_{j=1}^{n} p([\text{MASK}]_j = \phi_j(y) | T(x)).$$ (8)

3.3.2. Fine-Grained Evidence Aggregating Module

For the mentioned input, we feed the claim $\hat{c}$ and the retrieved evidence $\hat{e}_i \in \mathcal{E}$ into mBERT to achieve the corresponding representation. This is shown as follows:

$$h_{e_i} = \text{mBERT}(\hat{e}_i),$$

$$h_{c} = \text{mBERT}(\hat{c}).$$ (9)

Note that the embedding vectors of the claim and evidences are generated from the final hidden state of the special token [CLS]. Next, we extract the potential graph representations of the evidence nodes in the fine-gained evidence aggregating network, where information passing is guided by the noteworthy elements in the claim.
Semi-supervised graph networks have been used in the past to capture hidden information for a classification task [42,43]. Inspired by the previous studies, we propose a Recursive Graph Attention Network (RGAT) for evidence aggregating and reasoning.

In order to bridge information gaps between different pieces of evidence, a fully connected graph neural network is firstly constructed, where each network node represents one piece of evidence. In addition, every node is on a self-loop, and mainly considers the mining of fine-grained information and the flowing interaction of information from itself. Then, the three-layer recurrent network that deductively generates connected evidence chains is formulated for evidence reasoning. Specifically, the hidden states of nodes at layer \( l \) in the network are obtained as learning feature representations, as shown in Equation (10):

\[
h_l = \{h_{l1}, h_{l2}, \ldots, h_{lN}\},
\]

where \( h_{li} \in \mathbb{R}^{F \times 1} \) and \( F \) are the number of node features. The initial hidden state of each evidence node \( h_0^i \) is initialized by the evidence presentation:

\[
h_0^i = h_e^i.
\]

As far as the correlation coefficients between a node \( i \) and its neighbor \( j \) (\( j \in N_i \)) is concerned, we compute the coefficients as Equation (11):

\[
c_{ij} = \frac{1}{1 + \exp(-\mathbf{W}_l^{-1} \mathbf{h}_{l-1}^j \mathbf{h}_{l-1}^i)},
\]

where \( N_i \) denotes the neighbor set of node \( i \), \( \mathbf{W}_l^{-1} \in \mathbb{R}^{H \times 2F} \) and \( \mathbf{W}_l^1 \in \mathbb{R}^{1 \times H} \) are the weight matrix, and the operation \( \cdot \| \cdot \) indicates the concatenation of different variables.

Next, the coefficients \( c_{ij} \) are normalized with the softmax function, as shown in Equation (12):

\[
\alpha_{ij} = \text{softmax}_j(c_{ij}) = \frac{\exp(c_{ij})}{\sum_{k \in N_i} \exp(c_{ik})},
\]

Finally, through the recursive interactive fusion of evidence nodes layer by layer, the feature representation vector for node \( i \) at layer \( l \) is estimated as Equation (13):

\[
h_l^i = \sum_{j \in N_i} \alpha_{ij} h_{l-1}^j.
\]

Following this, the final hidden states of evidence nodes from stacking \( T \) layers are extracted as the component embedding representation of the evidence reasoning, which is seen in Equation (14):

\[
H_T = \text{CONCAT}(h_T^1, h_T^2, \ldots, h_T^N).
\]

As the next important step, we construct the evidence attention aggregator to gather the fine-grained information from the final embedding representation of the evidence reasoning. Specifically, what information is aggregated preferentially depends on how important it is to the claim, which is measured based on the multi-head attention mechanism [44], as shown in Equation (15):

\[
H^O = \text{MultiHead}(H_T, h_c) = \text{CONCAT}_{i \in [N_{head}]} \left[ \text{Head}^{(i)} \right] W_Q.
\]

In Equation (15), the \( \text{Head}^{(i)} \) is calculated by Equation (16):

\[
\text{Head}^{(i)} = \text{Attention}(H_T W_Q^{(i)}, h_c W_K^{(i)}, h_c W_V^{(i)}),
\]

where distinct parameter matrices \( W_Q^{(i)} \cdot W_K^{(i)} \in \mathbb{R}^{D_in \times d_k}, W_V^{(i)} \in \mathbb{R}^{D_in \times d_out} \) are learned for each head \( i \in [N_{head}] \) and the extra parameter matrix \( W_Q \in \mathbb{R}^{N_{head} \times D_out} \) projects the
concatenation of the $N_{\text{head}}$ head outputs (each in $\mathcal{R}^{d_{\text{out}}}$) to the output space $\mathcal{R}^{D_{\text{out}}}$. In the multi-head setting, we call $d_k$ the dimension of each head and $D_k = N_{\text{head}} d_k$ the total dimension of the query/key space.

Once the final state $H^O$ is obtained, we employ the multilayer perceptron to get the prediction $y_{\text{rgat}}$ in Equation (17):

$$y_{\text{rgat}} = \text{softmax}\left(\text{ReLU}\left(W^O H^O + b\right)\right), \quad (17)$$

where $W \in \mathcal{R}^{C \times D_{\text{out}}}$ and $b \in \mathcal{R}^{C \times 1}$ are parameters, and $C$ is the number of prediction labels.

Next, the cross-entropy is used as a loss function for optimizing the RGAT model, as shown in Equation (18):

$$L_{\text{rgat}}(y_{\text{rgat}}, y) = -\sum_{i=1}^{C} y^{(i)} \ast \log y_{\text{rgat}}^{(i)}. \quad (18)$$

### 3.4. Unified Affine Classification Layer

To combine the prompt-based claim verification module and the fine-grained evidence aggregation module uniformly for generating the final decision, we design the unified affine classification layer, which chiefly integrates the classification labels from the different module’s output. In other words, the last layer aims to strengthen the joint learning between multi-tasks for final efficiency and accurate predictions.

On the whole, the multi-task learning target is considered as the model optimizing problem concerning multiple objectives. Specifically, the naive approach to combining multi-objective losses would be to perform a weighted linear sum of the above losses for each module task, as shown in Equation (19):

$$L_{\text{total}} = L_1 + L_2 = \lambda_{\text{prompt}} L_{\text{prompt}} + \lambda_{\text{rgat}} L_{\text{rgat}}, \quad (19)$$

where the $L_1$ and $L_2$ are, respectively, the abstract expression of the loss function $L_{\text{prompt}}$ and $L_{\text{rgat}}$, $\lambda_{\text{prompt}} \in (0, 1)$, $\lambda_{\text{rgat}} \in (0, 1)$, and $\lambda_{\text{prompt}} + \lambda_{\text{rgat}} = 1$.

However, there are a series of issues employing the naive calculation method. Namely, model performance is extremely sensitive to weight values. These weight hyper-parameters are time-consuming to tune for the best performance, and are increasingly difficult with large models. Thus, we derive the unified loss function based on the previous work that proposed a principled approach to multi-task deep learning [45]—to learn optimal task weightings using ideas from probabilistic modeling based on the epistemic uncertainty and aleatoric uncertainty. The core calculation formula is shown in Equation (20):

$$L_{\text{total}}(W, \sigma_1, \sigma_2) = -\log p\left(y_{\text{promp}}, y_{\text{rgat}} \mid \hat{f}^W(x)\right) \propto \frac{1}{2\sigma_1^2} \left\| y_{\text{promp}} - \hat{f}^W(x) \right\|^2 + \frac{1}{2\sigma_2^2} \left\| y_{\text{rgat}} - \hat{f}^W(x) \right\|^2 + \log \sigma_1 \sigma_2$$

$$\approx \frac{1}{2\sigma_1^2} L_1(W) + \frac{1}{2\sigma_2^2} L_2(W) + \log \sigma_1 \sigma_2$$

$$= \frac{1}{2\sigma_1^2} L_1(W) + \frac{1}{2\sigma_2^2} L_2(W) + \log \sigma_1 + \log \sigma_2, \quad (20)$$

where $\sigma_1$ and $\sigma_2$ are the relative weight of the losses $L_1(W)$ and $L_2(W)$ adaptively, and the $\sigma_1$ and $\sigma_2$ are, respectively, the abstract expression of the weight parameters $\lambda_{\text{prompt}}$ and $\lambda_{\text{rgat}}$. On the other hand, in order to facilitate the modeling calculation, the $\frac{1}{2\sigma_1^2}$ and $\frac{1}{2\sigma_2^2}$ are, respectively, the approximate expression of the weight parameters $\lambda_{\text{prompt}}$ and $\lambda_{\text{rgat}}$.

More concretely, the final optimization objective can be seen as minimizing the loss function $L_{\text{total}}(W, \sigma_1, \sigma_2)$ with respect to $\sigma_1$ and $\sigma_2$ learning the weight of $W$ in the losses.
$L_1(W)$ and $L_2(W)$. As $\sigma_1$ indicates that the noise parameter for the label $y_{\text{prompt}}$ increases, the weight of $L_1(W)$ decreases. In other words, as the noise decreases, the weight of the respective objective increases. This last objective can be seen as learning the relative weights of the losses for each output. Large scale values $\sigma_1$ will decrease the contribution of $L_1(W)$, whereas small $\sigma_1$ will increase its contribution, and similarly for $\sigma_2$ of $L_2(W)$.

### 3.5. Procedure

In this section, we focus on revealing the model construction and optimization of Algorithm 1 in the training process, which not only includes the input and output data, but also the pre-processing and running process.

**Algorithm 1:** The process of training PEINet.

```plaintext
input : A sample set $X = \{x_1, x_2, \ldots, x_n\}$, where $x_i = \{\hat{c}_i, \hat{e}_1, \hat{e}_2, \ldots, \hat{e}_N\}$.
A corresponding label set $Y = \{y_1, y_2, \ldots, y_n\}$
output : The new model that will verify the multilingual claim.
1. Set Optimizer AdamW, and Initialize the iterator number iterNum, the batch size number batchNum, the weight parameters W, the learning rate lr, and other model parameters.;
2. trainSamples ← Preprocess(X,Y);
3. for i ← 0 to iterNum do
   4. batchSample ← Dataloader(trainSamples);
   5. for j ← 0 to batchNum do
      6. mBERTInput ← FeatureLoader(batchSample);
      7. mBERTOutput ← mBERT(mBERTInput);
      8. $M_{\text{Hidden}}$ ← mBERTOutput(Prompt\_MASK);
      9. PromptWord ← Hidden2Words($M_{\text{Hidden}}$);
     10. PromptProbs ← Words2Probs(PromptWord);
     11. PromptPredict ← Softmax(PromptProbs);
     12. $C_{\text{Hidden}}$ ← mBERTOutput(Claim\_CLS);
     13. $E_{\text{Hidden}}$ ← mBERTOutput(Evidence\_CLS);
     14. $E_{\text{Graph}}$ ← Rgat($E_{\text{Hidden}}$);
     15. FinalHidden ← Attention($C_{\text{Hidden}}, E_{\text{Graph}}$);
     16. RgatProbs ← FFN(FinalHidden);
     17. RgatPredict ← Softmax(RgatProbs);
     18. Calculate Prompt Cross Entropy Loss $L_{\text{prompt}}$;
     19. Calculate Rgat Cross Entropy Loss $L_{\text{rgat}}$;
     20. $L_{\text{total}}$ ← $\lambda_{\text{prompt}}L_{\text{prompt}} + \lambda_{\text{rgat}}L_{\text{rgat}}$;
     21. $L_{\text{total}}$ backward gradient;
     22. Update W lr;
   23. if save current model then
      24. Save model, tokenizer, optimizer;
      25. Evaluate Accuracy, F1, MeanLoss;
   26. else
      27. Evaluate Accuracy, F1, MeanLoss;
   28. end
   29. end
30. end
31. Return the best performance model in all evaluations.
```
For better intuitive understanding, the professional function namespace of the Pytorch framework is employed in the pseudocode. As a whole, the inference forecasting of the development and test process can be executed directly based on the model trained by the algorithm.

4. Experiments and Results

4.1. Datasets and Evaluation Metrics

We conduct our experiments on the large-scale benchmark dataset X-FACT, which includes 31,189 short claims labeled for veracity by expert fact-checkers and covers 25 typically diverse languages across 12 language families. According to the different rating scales for categorization, the label set is divided into seven labels by fact-checkers: True, Mostly True, Partly True, Mostly False, False, Complicated/Hard to Categorise, and Other. Table 1 shows the statistics of the dataset. Following the evaluation criteria of previous studies, we likewise report average accuracy, macro F1 scores, and standard deviations in the experience with different random seeds.

The dataset, shown in the Appendix A.1, provides a challenging multilingual fact-checking evaluation benchmark that measures in-domain classification, out-of-domain generalization, and zero-shot capabilities of the multilingual models. More specifically, the Test1 set contains fact-checks from the same languages and sources as the training set. Second, the out-of-domain test set (Test2) consists of claims from the same languages as the training set but are from a different fact-checking website source. The third test set (Test3) is the zero-shot set, which seeks to measure the cross-lingual transfer abilities of fact-checking systems. Test3 accommodates claims from the new language not contained in the training set. In general, the dataset is specially created to evaluate the various performances of the models that not only solve the difficulty of these multilingual fact-checking sets (Test1), but also fit both source-specific patterns (Test2) and language-specific patterns (Test3).

Table 1. X-Fact dataset details. X-FACT contains three challenges sets, namely, in-domain test (Test1), out-of-domain test (Test2), and zero-shot test (Test3). Each set consists of multilingual claims and pieces of evidence, whereby language numbers and pre-three languages are counted.

| Data Split          | Claims | Languages | Pre-3 Languages               |
|---------------------|--------|-----------|-------------------------------|
| Train               | 19,079 | 13        | Portuguese, Indonesian, Arabic|
| Development         | 2535   | 12        | Portuguese, Indonesian, Arabic|
| In-domain (Test1)   | 3826   | 12        | Portuguese, Indonesian, Arabic|
| Out-of-domain (Test2)| 2368  | 4         | Indonesian, Turkish, Portuguese|
| Zero-shot (Test3)   | 3381   | 12        | Albanian, Bengali, Russian    |

4.2. Setting

The core details of the model’s relevant parameters are provided for better evaluation and reproducibility.

Firstly, the hyper-parameter settings of the model architecture are as follows: In the shared encoding layer, the embedding layer of 50 dimensions is randomly initialized to encode the language family feature. For claim-evidence input pairs, all retrieved evidences related to the claim are inputted into the model, totaling five pieces. To maintain uniform input length, the input maximum sequence length is 360 due to constraints on the GPUs memory, where parts longer than the maximum length are truncated, and the shorter sequences are zero-padded. Next, we used the mBART-base model for all of our experiments. This model has 12 layers each with a hidden size of 768, the number of attention heads equal to 12, and the value for dropout in the hidden layer is 0.2. The total number of parameters in this model is 125 million. In the recursive graph attention network, we attempt to stack 0–3 recursive layers and analyze the effect of different layer numbers. Furthermore, the number of multi-attention heads is set to 0, 8, and 16, respectively, for evaluating the evidence aggregating performance. In the classification layer, the dynamic
weights are randomly initialized and gradually optimized in iteration. Furthermore, the class label number is 7, which is consistent with the number of task categories.

Next, regarding the training of the model, we optimize the model via the AdamW optimizer for the best performance, in which the initial learning rate is $1.95 \times 10^{-5}$, and the learning rate warm-up steps are set to 10% of the total training steps. In addition, the maximum batch size is 8 and the train epoch number is 10. Once each epoch training is completed, we evaluate the model performance and save the model once. All our models were run with four random seeds (seed = [1; 2; 3; 4]).

Lastly, for the experimental environment, all algorithm codes are developed based on the PyTorch framework, which is run on the NVIDIA GeForce RTX 3090 machine (24 GB VRAM) with a GPU accelerator. Furthermore, the average training time of our model is about 160 min. The experiments can be sped up with distributed data-parallel training via multiple GPUs on multiple computing machines.

4.3. Overall Performance

We compare our methods with the following several models, including top-performing models reported with the X-FACT dataset [9], and homogeneous models retrieved on the same dataset [42].

- **Majority.** The model directly generates classification results based on the majority label of the dataset distribution, i.e., Mostly False.
- **Claim-Only.** The model is only inputted with the textual claim, basically treating the task as a multi-class problem.
- **Attn-EA.** The procedure is emulated by developing an attention-based evidence aggregation model that operates on evidence documents retrieved from web searches with the claim using Google.
- **Graph-EA.** The framework for claim verification utilizes the BERT sentence encoder, the evidence reasoning network, and an evidence aggregator to encode, propagate, and aggregate information from multiple pieces of evidence.

**Common and Different Aspects.** In addition to the Majority model, which is calculated based on statistical methods, other models are designed based on the common architecture of a deep neural network, which leverages the mbert model to encode the claims and pieces of evidence. The core distinction between other models is the downstream network structure of the basic semantic encoder. Furthermore, the different network modules can be modeled to extract different features, which determines the final performance superiority of multilingual claim verification.

**Metadata Augmenting.** As far as metadata augmenting is concerned, to strengthen the model performance, the above models mainly adopt the scheme whereby the input is increased with additional metadata, and is represented as a sequence of the key-value form. In detail, the metadata includes the language, website name, claimant, claim date, and review date. If a certain field is not available for a claim, the value is set by none. Unlike the previous scheme, we only embed the metadata of language and language family for augmenting performances.

Table 2 reports the performance of our model and other models under the same evaluation conditions. Compared with the above models, PEINet has achieved state-of-the-art scores. Firstly, by vertically comparing model performance in training, validation, and testing, it can be seen that our model has no over-fitting. Next, by comparing the performance of the model on Test1 horizontally, the PEINet model outperforms the baseline models in the in-domain test, which demonstrates that the model has superior performance in multilingual automated fact-checking. Furthermore, our model significantly also outperforms previous methods not only on the Test2 set, but also on the Test3 set in further analysis, which means that our model has a better performance in both out-of-domain generalization, and zero-shot capabilities. The overall performance advantage of the model suggests that our motivation of the human fact-checking cognitive paradigm is valid and reasonable. In a separate analysis of the degree of performance superiority in three different test sets, our
model performs best in Test3, which further demonstrates the advantages of the model in few-/zero-shot scenarios.

Table 2. Average F1 scores (standard deviations) evaluated in the experiences.

| Set   | Splitting          | –Meta Data Augmentation | +Meta Data Augmentation |
|-------|--------------------|-------------------------|-------------------------|
|       | Majority           | Claim-Only              | Attn-EA                 | Graph-EA                 | Claim-Only | Attn-EA  | Graph-EA | Ours |
| Train | 7.0 (-)            | 38.0 (0.8)              | 40.2 (0.4)              | 38.9 (0.2)              | 39.8 (0.8) | 42.6 (1.1) | 40.1 (0.9) | 44.2 (1.2) |
| Dev   | 11.2 (-)           | 38.7 (0.6)              | 40.6 (0.7)              | 39.7 (0.5)              | 39.6 (0.6) | 42.0 (0.4) | 40.0 (0.4) | 42.9 (0.6) |
| Test1 | 6.9 (-)            | 38.2 (0.9)              | 38.9 (0.2)              | 38.0 (0.9)              | 39.4 (0.9) | 41.9 (1.2) | 39.2 (1.0) | 42.1 (0.7) |
| Test2 | 10.6 (-)           | 16.2 (0.9)              | 15.7 (0.1)              | 16.0 (0.9)              | 15.4 (0.8) | 15.4 (1.5) | 15.6 (1.2) | 17.3 (1.9) |
| Test3 | 7.6 (-)            | 14.7 (0.6)              | 16.5 (0.7)              | 13.7 (0.6)              | 16.7 (1.1) | 16.0 (0.3) | 15.0 (0.6) | 19.0 (1.1) |

4.4. Ablation Study

To investigate the effect of each component in PEINet, a concise summary of ablation analysis by removing different modules of our model is shown in Table 3. We employ ‘-UnifiedLossLayer’, ‘-FEAModule’, ‘-PCVModule’, and ‘-LFEncoderLayer’, which, respectively, represents: unified multi-task loss function, fine-grained evidence aggregating module, prompt-based claim verification module, and LF-shared encoding module. Individually, ‘-UnifiedLossLayer’ represents that the multi-task learning mechanism is not adopted, and both $\lambda_{\text{prompt}}$ and $\lambda_{\text{rgat}}$ are 0.5. Next, ‘-FEAModule’ means that the recursive graph attention network is replaced by the three-layer fully connected neural network. Then, ‘-PCVModule’ indicates that the prompt learning template is not employed, and the corresponding module is remodeled based on the claim embedding the special token [CLS]. Last, ‘-LFEncoder’ represents that the language family feature is not encoded, and only other meta-data features are adopted to strengthen the model.

The empirical results show that first, the removal of different components suffers different degrees of performance degradation. Second, the performance of the model without the fine-grained evidence aggregating module is subjected to a large reduction, which indicates that it is necessary to leverage the recursive graph attention between the claims and pieces of evidence to verify the claim, thus, contributing to boosting the final performance. Finally, without language family embedding, the model performance uniformly declined to a certain extent in each test scenario, which fully embodies the key role of language family metadata augmentation.

Table 3. Results of the ablation test of our PEINet on each test. For the sake of brevity, the corresponding acronym abbreviation is adopted to indicate modules or layers.

| Model             | Dev   | Test1  | Test2  | Test3  |
|-------------------|-------|--------|--------|--------|
| Ours              | 42.9  | 42.1   | 17.3   | 19.1   |
| -UnifiedLossLayer | 39.2  | 39.0   | 16.3   | 15.9   |
| -FEAModule        | 36.4  | 36.1   | 15.9   | 13.8   |
| -PCVModule        | 38.0  | 37.9   | 15.2   | 17.1   |
| -LFEncoder        | 35.9  | 39.1   | 17.4   | 17.8   |

4.5. Effectiveness of the Prompt Template

To verify the impact of different designed prompts, we manually design language-special and sequence-special contrast experiments. Therein, for language-special setting, except taking English as the prompt language, the Portuguese and Indonesian language are selected as representative languages, which belong to the top proportions of claim language. Furthermore, to evaluate whether there is an effective contribution of the language-special prompt, we also adopt the null prompt method, which simply concatenates inputs and the [MASK] tokens. As Table 4 shows, language-special prompts in correspondence with the source language of the claim do not perform best in our experiments, and the null prompt also limits the performance. On the contrary, we observe that the language of prompts close to the source language of the claim label can effectively boost the model performance.
The main reason is that on one head, the English prompt templates are better in line with the claim label. On the other head, errors may exist in the translated verbalizer. For a sequence-special setting, the experimental results suggest that template T1 in typical narration outperforms template T2 in a flashback.

Table 4. The Micro F1 scores evaluated with the different prompt templates encoded by different languages.

| Language-Special | Designed Template                                      | Dev  | Test1 | Test2 | Test3 |
|------------------|--------------------------------------------------------|------|-------|-------|-------|
| English          | T1(x) = The claim label may be [MASK][MASK]. x          | 42.9 | 42.1  | 17.3  | 19.1  |
|                  | T2(x) = Maybe [MASK][MASK]! Because of the claim that x  | 42.5 | 41.5  | 15.3  | 15.8  |
| Portuguese       | T1(x) = O rótulo da reivindicação pode ser [MÁSCARA][MÁSCARA]. x | 40.3 | 40.5  | 15.8  | 12.8  |
|                  | T2(x) = Talvez [MÁSCARA][MÁSCARA]! Por causa da alegação de que x | 40.1 | 40.4  | 16.0  | 12.9  |
| Indonesian       | T1(x) = Label klaim mungkin [MASK][MASK]. x             | 40.0 | 40.1  | 17.5  | 15.3  |
|                  | T2(x) = Mungkin [MASK][MASK]! Karena klaim bahwa x       | 40.4 | 40.2  | 17.6  | 16.3  |
| Null-Prompt      | T(x) = [MASK][MASK]. x                                 | 41.1 | 43.4  | 17.4  | 14.1  |

4.6. Effectiveness of RGAT Parameters Setting

Next, we focus on the effect of the different parameters in the RGAT module. In our observation, from the results in Table 5, the 8-head attention aggregator model with two recursive layers achieves the best results, which indicates that claims from the obscure and complicated subset require multi-step evidence chain propagation.

Table 5. The Micro F1 values on the different multi-head attention aggregators with different recursive layers (%).

| Recursive Layers | 8-Head Attention Aggregator | 16-Head Attention Aggregator |
|------------------|-----------------------------|------------------------------|
|                  | Dev  | Test1 | Test2 | Test3 | Dev  | Test1 | Test2 | Test3 |
| 0                | 36.4 | 36.1  | 15.9  | 13.8  | 36.4 | 36.1  | 15.9  | 13.8  |
| 1                | 42.6 | 39.1  | 16.9  | 18.1  | 36.5 | 37.5  | 16.6  | 19.0  |
| 2                | 42.9 | 42.1  | 17.3  | 19.1  | 40.4 | 38.9  | 16.7  | 15.5  |
| 3                | 42.8 | 38.0  | 15.2  | 17.1  | 40.6 | 36.6  | 16.0  | 15.5  |

We further analyze the results from different angles. From the perspective of model layer numbers, the performance of a model with three recursive layers is inferior to that of a model with two recursive layers, which suggests that the model is overfitting with the increase in layers. On the other head, when analyzing the parameters of the attention heads, the results show that the performance of the 16-head attention aggregators is overall lower than that of the 8-head models, which suggests that the multi-head attention mechanism recreates the role of information aggregation, and the redundancy errors are gradually introduced with the increase in multi-attention heads. Regarding the whole analysis, the results demonstrate the ability of our model to verify the claims with recursive excavation and aggregation of evidence.

4.7. Case Study and Error Analysis

In this section, we focus on the effect of error propagating based on experimental findings. We select representative samples from the train, dev, and test sets, which may reflect the prominent problem that PEINet confronts.

Through our analysis findings of the typical error cases, the retrieved evidence—a set of homogeneous snippets to classify the claim—may be the primary factor leading to the misjudgment of the model, which is exacerbated by space-limited abbreviation writing. For example, in Figure 3, to verify the claim “Wandering outside the home? Fire shoot in Malaysia with Drone.”, the pieces of evidence retrieved are almost indistinguishable—the #Evi1 and #Evi5 are almost similar, and the #Evi2, #Evi3, and #Evi4 are same. More critically, the evidence snippets are like replications of the claim, which cannot provide
sufficient information to classify the claim. This result may be explained by the fact that the new title of partial search snippets is mainly extracted as evidence, whereby information is insufficient. Next, through the evidence source link provided in the dataset, we further find the truth evidence from the full text of the evidence web pages. As shown in Figure 3, the core part of the evidence that determines the veracity of the claim is briefly listed in the table. A comparison of the retrieval evidence and truth evidence confirms that the claim verification component can not make the right inference with insufficient evidence.

**Claim:**

# Video “KELUYURAN DILUAR RUMAH? DI MALAYSIA DITEMBAK API PAKAI DRONE”
# Video “Wandering OUTSIDE THE HOME? FIRE SHOOT IN MALAYSIA WITH DRONE”

| Retrieved Evidence | Truth Evidence |
|---------------------|----------------|
| #Evi1: Oct 1, 2020 — Jakarta | EGINDO.co — BUKAN di Malaysia. Video itu adalah aksi peluncuran kembang api yang dilakukan oleh seorang influencer Brasil |
| #Evi2: Video “KELUYURAN DILUAR RUMAH? DI MALAYSIA DITEMBAK API PAKAI DRONE”, 22 Mar 2020. BUKAN di Malaysia. Video itu adalah aksi ...
| #Evi3: Video “KELUYURAN DILUAR RUMAH? DI MALAYSIA DITEMBAK API PAKAI DRONE” |
| #Evi4: Video “KELUYURAN DILUAR RUMAH? DI MALAYSIA DITEMBAK API PAKAI DRONE” |
| #Evi5: Mar 22, 2020 — BUKAN di Malaysia. Video itu adalah aksi peluncuran kembang api yang dilakukan oleh seorang influencer Brasil dengan teman-temannya di ...

| Truth Evidence |
|----------------|
| #Evi1: BUKAN di Malaysia. Video itu adalah aksi peluncuran kembang api yang dilakukan oleh seorang influencer Brasil dengan teman-temannya di Brazil, ...
| #Evi2: Faktanya, kejadian tersebut bukan terjadi di Malaysia, melainkan di Brasil, ...
| #Evi3: Akun Instagram Lucas Albert mengunggah video itu pada tanggal 14 Juli 2019 dengan narasi yang jika diterjemahkan sebagai berikut: “Teman-teman saya mengadakan acara barbekyu di jalan dan mereka tidak mengundang saya. Lihat apa yang saya lakukan dengan mereka.”...

| MetaData | Language Family: Austronesian; Source: cekfakta.com; Claimant: None; ClaimDate: 2020-03-22 |

**Figure 3.** A typical error analysis example in which the true label is misleading, while the predicted label is the other. The claim and retrieved evidence, which is highlighted in red font, come from the training dataset, and part of the truth evidence, which is marked with red font and is underlined, are manually searched from the evidence source website. For reference, translations are also shown. In addition, the metadata are also listed for integrity-checking.

To explore the effect of this issue, we test our models on an evidence-enhanced dev set, in which we add the multi-hop ground truth evidence chain using the full text of the
evidence pages. The experimental results show that augmenting evidence significantly improves model performance, which further is in agreement with the idea that fine-grained evidence aggregating and reasoning can contribute to verifying the claim. In addition, the error case analysis in this study also enhances an understanding that leveraging prompt learning to directly verify the claim possibly is a better idea for fact-checking the retrieval evidence to verify whether the claim is true or false or is similar to the claim. Hypothetically, this means it is worthwhile to design a better evidence retrieval pipeline, and the models are able to ingest large documents (web pages); thus, their performance increase could have been much more significant, which remains to be our focus in future research.

5. Discussion

In this section, we will discuss results by answering a series of research questions.

Do the different pre-trained multilingual models make wide differences? We emphatically analyze the performance of different pre-trained multilingual models employed in the framework. Table 6 shows that the model based on mBERT outperforms other similar models. With comparative analysis, the XLM-based model’s performance is close to that of the mBERT-based model, and the DistilBERT-based model’s overall performance is relatively poor, which is limited to performing direct zero-shot tasks due to the learning mechanism. The results suggest that similar large-scale pre-trained multilingual models can almost achieve the equivalent semantic understanding of the same claims’ input, so the core of fact-checking is still an evidence-aware inference.

Table 6. Performance analysis of different pre-trained multilingual models.

| Multilingual Model (BASE) | Dev  | Test1 | Test2 | Test3 |
|---------------------------|------|-------|-------|-------|
| DistilBERT                | 38.8 (0.7) | 36.1 (0.7) | 16.4 (1.8) | 17.5 (1.0) |
| XLM-Roberta               | 40.9 (0.6) | 39.9 (1.0) | 17.4 (1.8) | 17.8 (1.3) |
| mBERT                     | 42.9 (0.6) | 42.1 (0.7) | 17.3 (1.9) | 19.0 (1.1) |

What are the values of the learned relative weights of the losses in multi-task learning? We evaluate the performance of the PEINet model with different loss function calculation frameworks, where different methods correspond to different task weights. In the single-task learning framework, the task weights are set in advance. Under the multi-task learning framework, the relative weights of the losses are automatically generated through iterative training. The results of the contrastive analysis are presented in Table 7, which clearly illustrates the benefit of multi-task learning. Interestingly, what stands out in the table is that the task weight of the fine-grained evidence aggregation loss is higher than that of the prompt loss, which suggests that the verification of most multilingual claims would depend on the evidence chain excavation.

Table 7. Experiments of quantitative improvement with the multi-task loss. Results are shown from the three test sets.

| Loss Calculation               | Task Weights | Average Micro F1 Scores |
|--------------------------------|--------------|-------------------------|
|                               | \( \lambda_{prompt} \) | \( \lambda_{rgat} \) | Test1 | Test2 | Test3 |
| Prompt weight only             | 1.0          | 0.0                     | 36.0  | 15.2  | 13.9  |
| RGAT weight only               | 0.0          | 1.0                     | 38.0  | 15.8  | 17.2  |
| Unweighted sum of losses       | 0.5          | 0.5                     | 39.0  | 16.3  | 15.9  |
| Approx. optimal weights        | 0.11         | 0.89                    | 42.0  | 17.1  | 19.1  |

Can we improve performance with data augmentation of English corpus? A performance comparison is conducted by augmenting the dataset with 12,311 English claims and retrieved evidences from PolitiFact corpus. Average Micro F1 scores (and standard deviations) of the models are reported over four random runs in Table 8. The result that English data augmentation degrades the models’ performances is somewhat counter-intuitive.
A possible cause proposed by the dataset publishers is that augmenting data almost comes from the political field, and cross-domain tasks increase the difficulty of multilingual fact-checking, especially in the out-of-domain and zero-shot scene. Another significant factor in our observation is that the evaluation criteria between X-Fact dataset and PolitiFact dataset is inconsistent. However, with English data augmentation, different models have different levels of performance degradation, among which, the performance of PEINet is relatively more stable. The results further support the idea that language family embedding strengthens the model transfer learning in the zero-shot scene.

Table 8. Performance comparison analysis of different models with the augmenting data and original data.

| Model            | Test1   | Test2   | Test3   |
|------------------|---------|---------|---------|
| Claim-Only + Meta| 39.4 (0.9) | 15.4 (0.8) | 16.7 (1.1) |
| Attn-EA + Meta   | 41.9 (1.2) | 15.4 (1.5) | 16.0 (0.3) |
| PEINet           | 42.1 (0.7) | 17.3 (1.9) | 19.0 (1.1) |

6. Conclusions

In the paper, we propose a novel PEINet framework for multilingual fact-checking in the out-of-domain and zero-shot scenes. The framework endeavored to solve three challenges: truth or falsehood transfer learning of claims in cross-language and cross-domain, prompt-based verification without sufficient evidence, and fine-grained evidence aggregation and inference for fact verification. As alluded to in our review of the whole work, our model first focuses on the shared encoding mechanism of the language family metadata to strengthen the interactive learning of claims in different languages. Then, the prompt-based checking and synchronous evidence inference verification for the claim furthermore bridge gaps in multilingual fact-checking. Finally, the unified multi-task learning framework adjusts and optimizes the proportion weight of each module. The experiments achieve significant improvements, which proves that our framework performs well. Further analysis demonstrates the robustness and superiority of our model in cross-lingual and cross-domain fact verification tasks on two competitive test datasets. In the future, we will further explore: (1) How to leverage common sense knowledge or prior knowledge of relevant claims for breaking through the limitations of insufficient evidence to aggregate the vital clues. (2) Adaptive prompt template designs for multilingual claim verification to maybe play a key role in heuristic fact-checking in few-shot or zero-shot scenarios. (3) Interpretable claim verification is a potential research direction, which requires easy-to-interpret evidence inference and traceability.

Author Contributions: Conceptualization, X.L. and W.W.; methodology, X.L.; software, X.L. and W.W.; validation, X.L. and J.F.; formal analysis, X.L.; investigation, X.L.; resources, X.L. and J.F.; data curation, X.L., W.W. L.J. and C.L.; writing—original draft preparation, X.L.; writing—review and editing, X.L., W.W., J.F., L.J., H.K. and C.L.; visualization, X.L. and H.K.; supervision, L.J. and C.L.; project administration, X.L. and L.J.; funding acquisition, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (no. YZ35120378).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://github.com/LiXiaoyu0101/Multilingual-Fact-Checking (publicly accessed on 1 September 2022).
Conflicts of Interest: The authors declare that they do not have any conflicts of interest. This research does not involve any human or animal participation. All authors have checked and agreed with the submission.

Appendix A. Details on Dataset and Model

Appendix A.1. Dataset Construction

The composition of the X-Fact dataset is detailed in Table A1. There are 25 diverse languages across 12 language families among the dataset, which is divided into training dataset, development dataset, and three different task-oriented testing datasets. Number 0 indicates not including the language, and number 1 indicates including the language. Moreover, the language code and fact-checker are displayed accordingly.

Table A1. Details of the X-FACT dataset.

| ID | Language | Code   | FactChecker                 | Language Family | Train | Dev | Test1 | Test2 | Test3 |
|----|----------|--------|-----------------------------|-----------------|-------|-----|-------|-------|-------|
| 1  | Arabic   | ar     | misbar.com                  | Afro-Asiatic    | 1     | 1   | 0     | 0     | 0     |
| 2  | Bengali   | bn     | dailyo.in                   | IE:Indo-Aryan   | 0     | 0   | 0     | 0     | 1     |
| 3  | Spanish   | es     | chequeado.com               | IE:Romance      | 1     | 1   | 0     | 0     | 0     |
| 4  | Persian   | fa     | factnameh.com               | IE:Iranian      | 0     | 0   | 0     | 0     | 1     |
| 5  | Indonesian| id     | cekfakta.com                | Austronesian    | 1     | 1   | 1     | 0     | 0     |
| 6  | Indonesian| id     | cekfakta.tempo.co           | Austronesian    | 0     | 0   | 0     | 1     | 0     |
| 7  | Italian   | it     | pagellapolitica.it          | IE:Romance      | 1     | 1   | 1     | 0     | 0     |
| 8  | Italian   | it     | agi.it                      | IE:Romance      | 0     | 0   | 0     | 1     | 0     |
| 9  | Hindi     | hi     | aajtak.in                   | IE:Indo-Aryan   | 1     | 1   | 1     | 0     | 0     |
| 10 | Hindi     | hi     | hindi.newscrcher.in         | IE:Indo-Aryan   | 0     | 0   | 0     | 1     | 0     |
| 11 | Gujarati  | gu     | gujarati.newscrcher.in      | IE:Indo-Aryan   | 0     | 0   | 0     | 0     | 1     |
| 12 | Georgian  | ka     | factcheck.ge                | Kartvelian       | 1     | 1   | 1     | 0     | 0     |
| 13 | Marathi   | mr     | marathi.newscrcher.in       | IE:Indo-Aryan   | 0     | 0   | 0     | 0     | 1     |
| 14 | Punjabi   | pa     | punjabi.newscrcher.in.txt   | IE:Indo-Aryan   | 0     | 0   | 0     | 0     | 1     |
| 15 | Polish    | pl     | demagog.org.pl              | IE:Slavic       | 1     | 1   | 1     | 0     | 0     |
| 16 | Portuguese| pt     | piaiu.folha.uol.com.br      | IE:Romance      | 1     | 1   | 1     | 0     | 0     |
| 17 | Portuguese| pt     | poligrafo.sapo.pt           | IE:Romance      | 1     | 1   | 1     | 0     | 0     |
| 18 | Romanian  | ro     | factual.ro                  | IE:Romance      | 1     | 1   | 1     | 0     | 0     |
| 19 | Norwegian | no     | faktisk.no                  | IE:Germanic     | 0     | 0   | 0     | 0     | 1     |
| 20 | Sinhala   | si     | srilanka.factcrescendo.com  | IE              | 0     | 0   | 0     | 0     | 1     |
| 21 | Serbian   | sr     | istinomer.rs                | IE:Slavic       | 1     | 1   | 1     | 0     | 0     |
| 22 | Tamil     | ta     | youturn.in                  | Dravidian       | 1     | 1   | 1     | 0     | 0     |
| 23 | Albanian  | sq     | kalixo.com                  | IE:Albanian     | 0     | 0   | 0     | 0     | 1     |
| 24 | Albanian  | sq     | faktoje.al                  | IE:Albanian     | 0     | 0   | 0     | 0     | 1     |
| 25 | Russian   | ru     | factcheck.kz                | IE:Slavic       | 0     | 0   | 0     | 0     | 1     |
| 26 | Turkish   | tr     | dogrulukpayi.com            | Turkic          | 1     | 1   | 1     | 0     | 0     |
| 27 | Turkish   | tr     | teyit.org                   | Turkic          | 0     | 0   | 0     | 1     | 0     |
| 28 | Azerbaijani| az     | faktyoxda.info              | Turkic          | 0     | 0   | 0     | 0     | 1     |
| 29 | Portuguese| pt     | aosfatos.org                | IE:Romance      | 0     | 0   | 0     | 0     | 1     |
| 30 | German    | de     | correctiv.org               | IE:Germanic     | 1     | 1   | 1     | 0     | 0     |
| 31 | Dutch     | nl     | nieuwscheckers.nl           | IE:Germanic     | 0     | 0   | 0     | 0     | 1     |
| 32 | French    | fr     | frafricacheck.org           | IE:Romance      | 0     | 0   | 0     | 0     | 1     |

Appendix A.2. Model Run Times

The compared baseline models are specified, of which the average running times trained in the same conditions and average inference times tested in the same conditions are presented in Table A2. Further analysis suggests that the average time complexity of our framework is at the same level as that of other representatives. Thus, the experiment demonstrates that the improvement of our model performance depends on the design of modeling architecture rather than the accumulation of complex networks.
Table A2. The statistical results of the average training runtime in per epoch and the inference runtime in per sample. For short, minute is min, and millisecond is ms.

| DataSet        | Average Training Runtime (per Epoch) | Average Inference Runtime (per Sample) |
|----------------|-------------------------------------|---------------------------------------|
|                | Claim-Only | Attn-EA | PEINet | Claim-Only | Attn-EA | PEINet |
| X-FACT         | 13 min     | 20 min  | 17 min | 100 ms     | 118 ms  | 117 ms  |
| X-FACT-ENGLISH | 19 min     | 32 min  | 28 min | 100 ms     | 118 ms  | 117 ms  |

References

1. Allen, J.; Howland, B.; Mobius, M.; Rothschild, D.; Watts, D.J. Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv.* 2020, 6, eaay3539. [CrossRef] [PubMed]

2. Islam, M.S.; Sarkar, T.; Khan, S.H.; Kamal, A.H.M.; Hasan, S.M.; Kabir, A.; Yeasmin, D.; Islam, M.A.; Chowdhury, K.I.A.; Anwar, K.S.; et al. COVID-19–related infodemic and its impact on public health: A global social media analysis. *Am. J. Trop. Med. Hyg.* 2020, 103, 1621. [CrossRef] [PubMed]

3. Kazemi, A.; Garimella, K.; Gaffney, D.; Hale, S.A. Claim matching beyond English to scale global fact-checking. *arXiv* 2021, arXiv:2106.00853.

4. Schwarz, S.; Theophilo, A.; Rocha, A. Emet: Embeddings from multi-lingual-encoder transformer for fake news detection. In *Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 4–8 May 2020; pp. 2777–2781.

5. Shahi, G.K.; Nandini, D. FakeCovid–A multilingual cross-domain fact check news dataset for COVID-19. *arXiv* 2020, arXiv:2006.11343.

6. Roy, A.; Ekbal, A. MulCoB-MulFaV: Multi-modal Content Based Multilingual Fact Verification. In *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, 18–22 July 2021; pp. 1–8.

7. Martin, A.; Huertas-Tato, J.; Huertas-García, Á.; Villar-Rodríguez, G.; Camacho, D. FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowl.-Based Syst.* 2022, 251, 109265. [CrossRef]

8. Lee, N.; Bang, Y.; Madotto, A.; Fung, P. Towards Few-shot Fact-Checking via Perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, 6–11 June 2021; pp. 1971–1981. [CrossRef]

9. Gupta, A.; Srikumar, V. X-Fact: A New Benchmark Dataset for Multilingual Fact Checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, Bangkok, Thailand, 1–6 August 2021. [CrossRef]

10. Kotonya, N.; Toni, F. Explainable automated fact-checking: A survey. *arXiv* 2020, arXiv:2011.03870.

11. Guo, Z.; Schlichtkrull, M.; Vlachos, A. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguist.* 2022, 10, 178–206. [CrossRef]

12. Lowrey, W. The emergence and development of news fact-checking sites: Institutional logics and population ecology. *J. Stud. 2017, 18, 376–394. [CrossRef]*

13. Niewiński, P.; Pszona, M.; Janicka, M. GEM: Generative enhanced model for adversarial attacks. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, Hong Kong, China, 3–7 November 2019; pp. 20–26.

14. Wang, W.Y. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv* 2017, arXiv:1705.00648.

15. Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. Fever: A large-scale dataset for fact extraction and verification. *arXiv* 2018, arXiv:1803.05355.

16. Thorne, J.; Vlachos, A. Adversarial attacks against fact extraction and verification. *arXiv* 2019, arXiv:1903.05543.

17. Aly, R.; Guo, Z.; Schlichtkrull, M.; Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv* 2021, arXiv:2106.05707.

18. Zeng, X.; Abumansour, A.S.; Zubíaga, A. Automated fact-checking: A survey. *Lang. Linguist. Compass* 2021, 15, e12438. [CrossRef]

19. Hanselovská, A.; Zhang, H.; Li, Z.; Sorokin, D.; Schiller, B.; Schulz, C.; Gurevych, I. Ukp-athene: Multi-sentence textual entailment for claim verification. *arXiv* 2018, arXiv:1809.01479.

20. Nie, Y.; Chen, H.; Bansal, M. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6859–6866.

21. Wu, L.; Rao, Y.; Sun, L.; He, W. Evidence inference networks for interpretable claim verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Virtually, 2–9 February 2021; Volume 35, pp. 14058–14066.

22. Hassan, N.; Arslan, F.; Li, C.; Tremayne, M. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, NS, Canada, 13–17 August 2017; pp. 1803–1812.

23. Baly, R.; Mohtarami, M.; Glass, J.; Márquez, L.; Moschitti, A.; Nakov, P. Integrating stance detection and fact checking in a unified corpus. *arXiv* 2018, arXiv:1804.08012.
24. Khouja, J. Stance Prediction and Claim Verification: An Arabic Perspective. In Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Online, 9–10 July 2020; pp. 8–17. [CrossRef]
25. Nørregaard, J.; Derczynski, L. DANFEVER: Claim verification dataset for Danish. In Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), Reykjavik, Iceland, 31 May–2 June 2021; pp. 422–428.
26. Vogel, I.; Meghana, M. Detecting fake news spreaders on twitter from a multilingual perspective. In Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, Australia, 6–9 October 2020; pp. 599–606.
27. Patwa, P.; Sharma, S.; Pykl, S.; Gupta, V.; Kumari, G.; Akhtar, M.S.; Ekbal, A.; Das, A.; Chakraborty, T. Fighting an infodemic: COVID-19 fake news dataset. In Proceedings of the International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, Virtual Event, 8 February 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 21–29.
28. Mattern, J.; Qiao, Y.; Kerz, E.; Wiechmann, D.; Strohmaier, M. FANG-COVID: A New Large-Scale Benchmark Dataset for Fake News Detection in German. In Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7–11 November 2021; pp. 78–91. [CrossRef]
29. Alhindi, T.; Alabdulkarim, A.; Alshehri, A.; Abdul-Mageed, M.; Nakov, P. Arastance: A multi-country and multi-domain dataset of arabic stance detection for fact checking. arXiv 2021, arXiv:2104.13559.
30. Nielsen, D.S.; McConville, R. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. arXiv 2022, arXiv:2202.11684.
31. Dementieva, D.; Panchenko, A. Fake news detection using multilingual evidence. In Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, Australia, 6–9 October 2020; pp. 775–776.
32. Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A.H.; Riedel, S. Language models as knowledge bases? arXiv 2019, arXiv:1909.01066.
33. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 2020, 33, 1877–1901.
34. Lin, X.V.; Mihaylov, T.; Artetxe, M.; Wang, T.; Chen, S.; Simig, D.; Ott, M.; Goyal, N.; Bhosale, S.; Du, J.; et al. Few-shot Learning with Multilingual Language Models. arXiv 2021, arXiv:2112.10668.
35. Floridi, L.; Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. Minds Mach. 2020, 30, 681–694. [CrossRef]
36. Pan, L.; Chen, W.; Xiong, W.; Kan, M.Y.; Wang, W.Y. 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), Virtual Event, 6–11 June 2021; Association for Computational Linguistics: Cambridge, MA, USA; pp. 125–129.
37. Pires, T.; Schlinger, E.; Garrette, D. How multilingual is multilingual BERT? arXiv 2019, arXiv:1906.01502.
38. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv 2021, arXiv:2107.13586.
39. Lv, B.; Jin, L.; Zhang, Y.; Wang, H.; Li, X.; Guo, Z. Commonsense Knowledge-Aware Prompt Tuning for Few-Shot NOTA Relation Classification. Appl. Sci. 2022, 12, 2185. [CrossRef]
40. Han, X.; Zhao, W.; Ding, N.; Liu, Z.; Sun, M. Ptr: Prompt tuning with rules for text classification. arXiv 2021, arXiv:2105.11259.
41. Zhou, J.; Han, X.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. GEAR: Graph-based evidence aggregating and reasoning for fact verification. arXiv 2019, arXiv:1908.01843.
42. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. arXiv 2016, arXiv:1609.02907.
43. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
44. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7482–7491.