MAGIC: Multi-Argument Generation with Self-Refinement for Domain Generalization in Automatic Fact-Checking

Wei-Yu Kao, An-Zi Yen
Department of Computer Science, National Yang Ming Chiao Tung University
No. 1001, Daxue Rd. East Dist., Hsinchu City 300093, Taiwan
{wayner.cs09, azyen}@nycu.edu.tw

Abstract

Numerous studies have been conducted on automatic fact-checking, driven by its importance in real-world applications. However, two challenges persist: (1) extracting pivotal evidence from extensive documents, and (2) verifying claims across diverse domains. On one hand, current retrieval methods are limited in their ability to concisely retrieve evidence, which results in poor performance. On the other hand, retrieved evidence derived from different sources strains the generalization capabilities of classifiers. This paper explores the task of cross-domain fact-checking and presents the XClaimCheck dataset, which consists of claims from multiple domains. We propose a framework featuring a multi-argument generation technique. We leverage multi-argument generation to reconstruct concise evidence from large amounts of evidence retrieved from different sources. In addition, a self-refinement mechanism is introduced to confirm that the generated arguments are consistent with the content of the evidence. Experimental results show that our proposed framework is effective in identifying the veracity of out-of-domain claims, particularly those that are partially true or false.

Keywords: Automatic Fact-Checking, Domain Generalization, Multi-Argument Generation, Self-Refinement

1. Introduction

In this era of information overload, from social media feeds to online news outlets, people are bombarded by an overwhelming amount of information. This deluge of data, while beneficial in many respects, has also given rise to significant challenges. The most prominent among these is the urgent need for automatic fact-checking. As the volume of information continues to grow, so does the difficulty of discerning fact from fiction. Information that people receive, particularly through online platforms, significantly influences their mindset and perceptions. It shapes their impression of public figures, sways their opinions on critical issues, and can even impact their decision-making processes. Hence numerous studies have explored methodologies for automatic fact-checking.

Previous work can be divided into evidence-less methods and those that make use of selected evidence. Early approaches focused primarily on the claims themselves, attempting to predict their veracity based on various claim-related characteristics. For example, Popat et al. (2017) and Rashkin et al. (2017) conduct analyses of linguistic attributes in untrustworthy text. Others (Zubiaga et al., 2016; Derczynski et al., 2017; Wang, 2017; Fajcik et al., 2019; Li et al., 2019; Atanasova et al., 2019; Gorrell et al., 2019) make use of meta-information, examining factors such as the claimant's identity or public reactions to statements. However, such approaches are limited by their reliance on relatively scarce information.

Determining the truthfulness of a claim without additional information is a challenging task, primarily due to the lack of obvious veracity and the often contentious nature of such claims. Even professional fact-checkers require specific evidence to either substantiate or debunk claims. Thus some studies focus on evidence-based automatic fact-checking methods. For instance, Thorne et al. (2018) employ sparse vectors like TF-IDF to retrieve relevant documents. The utilization of both sparse vectors and the BERT-based dense passage retriever (DPR) (Karpukhin et al., 2020) has also been explored (Jiang et al., 2020; Park et al., 2022; Jiang et al., 2021; Stammbach, 2021; Khan et al., 2022). Casillas et al. (2022) and Fajcik et al. (2022) propose novel frameworks for concatenating claims and evidence embeddings.

Despite notable progress in the field, two significant challenges persist. The first challenge revolves around the extraction of pivotal evidence from extensive documents. Pan et al. (2023b) investigate evidence granularity influencing the generalizability of fact-checking models. Sentence-level evidence consists of carefully selected fine-grained information, while document-level evidence carries coarse-grained content. Compared to sentence-level evidence, document-level evidence makes fact-checking models face greater difficulty. This may be because document-level evidence requires more advanced reasoning skills; indeed, models are typically subject to strict context window limitations. Current approaches such as the dense passage retriever struggle to organize all relevant information concisely and comprehensively for pre-

LREC-COLING 2024, pages 10891–10902

diction models. This limitation can lead to the omission of crucial evidence, thereby impacting the correctness of the fact-checking results. In real-world scenarios, it is crucial to thoroughly assess a range of credible data sources in order to extract valuable information. This further emphasizes the importance of effective evidence reconstruction.

The second challenge arises from the proliferation of data across diverse domains and the rapid dissemination of information. Present methods tend to overlook domain generalization and rely heavily on models fine-tuned for specialized domains. While Pan et al. (2023b) propose two approaches to tackle the issue, their methods primarily emphasize the generation of additional data and the pretraining of domain-specific models. They acknowledge that limited progress has been achieved in addressing these challenges.

To address the aforementioned challenges, we propose MAGIC (**m**ulti-**a**rgument **g**eneration to reconstruct retrieved evidence for use **i**n fact-**c**hecking), a pilot framework. For each document, we retrieve evidence and then employ a language model to generate a corresponding argument based on the retrieved evidence. Given the impressive capabilities of LLMs in semantic understanding and sentence generation, we incorporate an LLM into our framework as the generator. This approach efficiently reconstruct evidence retrieved from extensive documents into concise and salient arguments. In addition, we use a self-refinement mechanism to confirm that the generated argument faithfully represents the perspective derived from the retrieved evidence.

To evaluate the efficacy of the proposed method in cross-domain fact-checking, it is crucial to have claims from various domains. Many studies have constructed datasets for fact-checking research covering topics such as healthcare (Kotonya and Toni, 2020), political issues (Wang, 2017), and multimodal fake news originating from social media platforms (Nakamura et al., 2020). Among these datasets, WatClaimCheck (Khan et al., 2022) covers claims from various fact-checking websites, accompanied by their related review articles, premise articles, and claim verdicts. However, the claims in WatClaimCheck are not divided into specific domains. Therefore, we extend WatClaimCheck, selecting claims from PolitiFact,[1] and annotate them based on the website's categorization, resulting in 26 distinct topics. Given that some domains have related themes, we further manually group them into five domains, including "Public Policy and Finance", "Political Issues", "Legal and Regulatory Affairs", "Infrastructure and Services", and "Global Affairs and Security". The details will be described in the following section. In sum, our contributions can be summarized as follows:

- We introduce the task of cross-domain fact-checking, which addresses the challenges of domain adaptation to retrieve evidence effectively from premise articles and determine the veracity of claims.
- We present the XClaimCheck dataset,[2] annotated with domain information and further categorized into five major domains, establishing a new benchmark for cross-domain fact-checking.
- We propose a framework which employs multi-argument generation with self-refinement in fact-checking (MAGIC). Experimental results demonstrate promising performance in the out-of-domain claim verification, especially in determining the veracity of partially true/false claims. Moreover, self-refinement improves both in-domain and out-of-domain fact-checking.

2. Related Work

2.1. Multi-Domain Fact-Checking

Several datasets have been introduced for cross-domain fact-checking. These datasets primarily include data from a wide range of sources, including Wikipedia (Thorne et al., 2018), fact-checking websites (Khan et al., 2022), news portals (Kotonya and Toni, 2020), and social media platforms (Nakamura et al., 2020). Among these works, Augenstein et al. (2019) has to date constructed the most comprehensive claim and evidence data, sourcing from 26 different websites. Pan et al. (2023b) scrutinize 11 fact-checking datasets across six domains, incorporating compelling topics such as climate, science, and health. However, domain generalization in automatic fact-checking is yet to be explored. Some have investigated the adaptation of misinformation detection methods to unseen domains such as COVID-19 (Yue et al., 2022) and scientific claims (Wadden et al., 2022, 2020). In this paper, we conduct a pilot experiment on domain generalization for fact-checking, covering a broader range of domains, to provide a holistic understanding and analysis of the challenges.

2.2. Application of LLM Methods in Fact-Checking

Large language model (LLM) methods have been widely employed in summarization, translation, and question-answering tasks (Brown et al., 2020; Goyal et al., 2022; McCarthy et al., 2022). These

[1] https://www.politifact.com/

[2] https://github.com/NYCU-NLP-Lab/XClaimCheck

Public policy and finance		Political issues		Legal and regulatory affairs		Infra. and services		Global affairs and security	
Topic	Count	Topic	Count	Topic	Count	Topic	Count	Topic	Count
Federal budget	824	Elections	1,167	Legal issues	554	Technology	145	Foreign policy	693
State budget	734	Candidate bio	801	LGBTQ	138	Energy	448	Immigration	983
Taxes	1,242	Jobs	914	Criminal justice	456	Transportation	267	Religion	235
Economy	1,337	Govt. regulation	248	Social sec.	168	Education	946	History	589
Health care	1,573			Homeland sec.	307	Sports	142	Military	420
Environment	436							Terrorism	410
Sum	6,146	Sum	3,130	Sum	1,623	Sum	1,948	Sum	3,330

Table 1: Domains and corresponding topics

methods have demonstrated significant enhancements in performance. Given the limited exploration of their potential utility in the domain of fact-checking, there is a highly promising avenue for further investigation in this area. Pan et al. (2023a) introduce a novel framework that integrates LLM methods into fact-checking tasks. Their methodology involves the use of few-shot learning to generate a reasoning process that verifies each component of the claim. They employ models such as BERT (Devlin et al., 2019) or FLAN-T5 (Chung et al., 2022) to construct subtask functions that facilitate the reasoning process. For subtasks such as fact-verification and prediction, their methods involve straightforward one-stage prediction, a process which involves concatenating all textual inputs into a single sequence, which is then used as input for sequence-to-sequence models to generate answers. Consequently, their work focuses on reasoning which portions of the claim need verification. In this work, by contrast, we seek to develop techniques that enhance evidence reconstruction and domain generalization for fact verification.

3. Dataset Construction

WatClaimCheck consists of 33,721 claims collected from eight fact-checking websites. The premise articles and review articles are valuable for developing claim inference methods. Hence, in this work, we extend WatClaimCheck to study cross-domain fact-checking. Given the necessity of associating claims with topic labels, we chose PolitiFact as our data source due to its stable and diverse collection of domains. We thus collected a total of 15,867 claims from PolitiFact in WatClaimCheck to construct XClaimCheck.

As the claims in WatClaimCheck were not tagged with topic labels, we obtained the topic information from PolitiFact. Table 1 shows the statistics of our XClaimCheck dataset. We identified a total of 26 representative topics, each containing a substantial amount of data suitable for our fact-checking task. Note that certain claims, due to their multifaceted nature, pertain to multiple topics. In such cases, we assigned these claims and their associated data to all relevant topics, resulting in a total of 16,177 instances. Then we organized these 26 topics into five distinct domains. Each topic was grouped with other relevant topics whenever possible. The five domains in XClaimCheck are “Public Policy and Finance”, “Political Issues”, “Legal and Regulatory Affairs”, “Infrastructure and Services”, and “Global Affairs and Security”, respectively.

	False	Mostly False	Half True	Mostly True	True
Training	3,064	1,827	1,916	1,782	1,140
Validation	987	625	668	541	403
Test	1,022	576	651	569	406
Sum	5,073	3,028	3,235	2,892	1,949

Table 2: Rating count distribution

Table 2 shows the rating distribution in our dataset. PolitiFact’s claim ratings have six distinct labels, but we additionally assigned the label “Pants on Fire” to claims that are categorically “False”, because “Pants on Fire” denotes claims that are egregiously incorrect; it essentially serves as a more severe form of a “False” claim. Consequently, our dataset was ultimately labeled with five claim ratings. To prevent overlap, we started off by dividing all instances into distinct training, validation, and test sets.

4. Methodology

In this section, we introduce MAGIC, our fact-checking framework. The modules in MAGIC are an evidence retriever, a multi-argument generator, an alignment verifier, and a fact-checker. Figure 1 is an overview of MAGIC. Let C_i denote the i-th claim, and let the set of premise articles associated with C_i be $P_i = \{p_{i,1}, p_{i,2}, \ldots, p_{i,N}\}$, where $i \in [1, I]$. I is the total number of claims and N is the number of premise articles for the i-th claim. MAGIC first generates multiple arguments $\mathcal{A}_i = \{A_{i,1}, A_{i,2}, \ldots, A_{i,N}\}$ for the i-th claim, with each argument $A_{i,n}$ corresponding to the n-th premise article. Next, each $A_{i,n}$ is refined by our self-refinement mechanism, resulting in a refined set of arguments $\mathcal{A}_i^{'}$. Finally, the fact-checker predicts the veracity of C_i based on the content of C_i and $\mathcal{A}_i^{'}$.

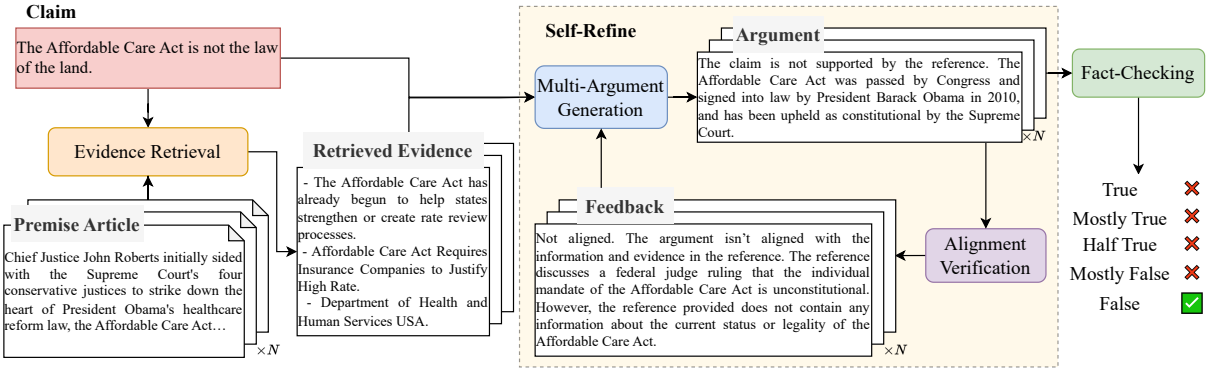

Figure 1: MAGIC overview

4.1. Evidence Retrieval

The n-th premise article $p_{i,n}$ of the i-th claim contains sentences $p_{i,n} = \{ps^n_{i,1}, ps^n_{i,2}, \ldots, ps^n_{i,M}\}$, where M is the number of the sentences within $p_{i,n}$. Our goal is to retrieve the sentences from $p_{i,n}$ that serve as evidence related to C_i. Hence, we employ the dense passage retriever (DPR) proposed by Karpukhin et al. (2020). To train the DPR model, we follow the method proposed by Khan et al. (2022), which uses expert-written review articles as the ground truth.

C_i corresponds to review article R_i, which contains L sentences $\{rs^1_i, rs^2_i, \ldots, rs^L_i\}$. Every sentence within R_i is used as positive example for training. In addition, DPR training requires negative and hard negative examples, the latter being instances closely resembling positive examples. For a given claim C_i, we use the sentences in R_j from the j-th claim C_j that share the same domain as C_i but belong to a different topic to serve as the negative examples of C_i. Those sentences in R_k from C_k that share the same topic as C_i are selected as the hard negative examples.

Let $R^+_i = \{rs^+_{i,1}, \ldots, rs^+_{i,\alpha}\}$ denote the set of positive examples of C_i. $R^-_i = \{rs^-_{j,1}, \ldots, rs^-_{j,\beta}\}$ and $R^\star_i = \{rs^\star_{k,1}, \ldots, rs^\star_{k,\gamma}\}$ are the sets of negative examples and hard negative examples, respectively. The training data is $\{\langle C_i, R^+_i, R^-_i, R^\star_i\rangle\}^d_{i=1}$, which contains d instances. Each instance consists of one claim C_i, α positive examples $rs^+_{i,q}$, β negative examples $rs^-_{j,r}$, and γ hard negative examples $rs^\star_{k,s}$. We optimize the loss function based on the negative log likelihood of $rs^+_{i,q}$:

$$L = -\log \frac{S(C_i,+)}{\sum^{\alpha}_{q=1} S(C_i,+) + \sum^{\beta}_{r=1} S(C_i,-) + \sum^{\gamma}_{s=1} S(C_i,\star)} \quad (1)$$

where $S(C_i,+)$, $S(C_i,-)$, and $S(C_i,\star)$ are defined as $e^{\mathrm{sim}(C_i, rs^+_{i,q})}$, $e^{\mathrm{sim}(C_i, rs^-_{j,r})}$, and $e^{\mathrm{sim}(C_i, rs^\star_{k,s})}$, respectively. $\mathrm{sim}(v,u)$ is a function that computes the dot product between claim vector v and sentence vector u.

Since there are five domains in this work, we train a separate retriever for each domain to mimic the real-world scenario where the retrieval model is not pre-trained on other domains.

4.2. Multi-Argument Generation

Using the evidence retriever, we retrieve sentences relevant to C_i from a vast number of premise articles. For each premise article, we select the top 50 sentences based on the ranking scores as the retrieved evidence. However, this yields a substantial number of sentences, and these sentences are disorganized and not coherent. To obtain concise and insightful evidence, we construct a generator to reconstruct the retrieved evidence and produce multiple arguments based on various evidence sources. We refer to this process as multi-argument generation. Apart from reconstructing the evidence, multi-argument generation also serves a crucial role in filtering out irrelevant information. In essence, evidence retrieval can include irrelevant or noisy results, making it critical to exclude such information. In multi-argument generation, given a claim C_i, its retrieved evidence $\mathcal{E}_{i,n}$, and the generation prompt $\mathcal{P}_{gen}$, we instruct the LLM $\mathcal{M}$ to assess whether the input evidence directly addresses the claim. If the model responds with "not relevant", triggering the condition $\mathrm{drop}(\cdot)$, this evidence is disregarded and not considered in the final output y_i. Arguments generated from all retrieved evidence are aggregated into $\mathcal{A}_i$. In sum, the process is formulated as

$$A_{i,n} = \mathcal{M}(C_i, \mathcal{E}_{i,n}; \mathcal{P}_{gen}) \quad (2)$$

$$\mathcal{A}_i = \{\langle A_{i,n}\rangle, \text{if}(\neg\mathrm{drop}(A_{i,n}))\}^N_{n=1} \quad (3)$$

In contrast to summarization, in this framework, argument generation generates perspectives based on evidence to assess the veracity of claims. Additionally, the LLM-based multi-argument generation offers a concise perspective that clarifies the interpretation of the claim based on the provided evidence.

4.3. Self-Refinement

In the process of multi-argument generation, LLMs sometimes fail to generate arguments that are aligned with the evidence received. This may be because feeding an LLM with claims that are subjective and potentially misleading can cause the model to producing biased assertions in favor of those claims. Such outputs might not genuinely take into account the relevant evidence, and instead be misled by the content of the claim. Inspired by the method proposed by Madaan et al. (2023), which utilizes the model itself to generate feedback and correct the previous output, we add self-refinement to MAGIC to verify that the generated argument is aligned with the corresponding evidence. That is, the generated argument is directly derived from the evidence. In this way, we ensure that the veracity prediction of the claim is grounded in arguments drawn from the evidence.

Algorithm 1 shows how the self-refinement mechanism is integrated into MAGIC, working with the multi-argument generator to iteratively refine the argument. Formally, we obtain the argument at the t-th turn $A^t_{i,n}$ generated by $\mathcal{M}$ with the argument generation prompt $\mathcal{P}_{gen}$, then we input $A^t_{i,n}$ to $\mathcal{M}$ so that $\mathcal{M}$ can refine $A^t_{i,n}$ based on $\mathcal{E}_{i,n}$. We obtain the feedback $f^t_{i,n} = \mathcal{M}(A^t_{i,n}, \mathcal{E}_{i,n}; \mathcal{P}_f)$, where $\mathcal{P}_f$ is the feedback prompt, yielding $f^t_{i,n}$, the model's verification of whether the generated argument aligns with the retrieved evidence. If not aligned, we obtain a new generated argument $A^{t+1}_{i,n}$ from $\mathcal{M}$ with the self-refinement prompt $\mathcal{P}_{rf}$.

The above process is repeated T times ($1 \leq t \leq T$). When $t = 1$, given a claim C_i and evidence $\mathcal{E}_{i,n}$ retrieved from the n-th premise article, we obtain the initial argument $A^0_{i,n} = \mathcal{M}(C_i, \mathcal{E}_{i,n}; \mathcal{P}_{gen})$. At other turns, we also input $f^t_{i,n}$ to $\mathcal{M}$ to refine the argument, i.e., $A^t_{i,n} = \mathcal{M}(C_i, \mathcal{E}_{i,n}, f^t_{i,n}; \mathcal{P}_{rf})$. In this work, we set T to 10. In $\mathcal{P}_f$, we instruct the model to output "not aligned" if the argument and evidence do not align, and "aligned" otherwise. Hence, if $\mathcal{M}$ suggests "aligned", the self-refinement process terminates. We construct a stop function $\mathrm{stop}_{align}(\cdot)$ to identify whether the process should be stopped. The input of $\mathrm{stop}_{align}(\cdot)$ is $f^t_{i,n}$. We set a maximum number of trials, $T = 10$, to prevent $\mathcal{M}$ from continuously failing to produce either "aligned" or "not aligned" outputs, and avoid infinite loops.

Note that the claim is not provided as part of the input to avoid mirroring the multi-argument generation process and to prevent the model from being misled by the claim itself. By having the model compare the alignment of two textual contents, the complexity of the task is reduced and the effectiveness of the self-refinement process is enhanced.

Algorithm 1 Self-refinement algorithm for multi-argument generation

Require: Claim C_i, evidence $\mathcal{E}_{i,n}$, model $\mathcal{M}$, prompts {$\mathcal{P}_{gen}$,$\mathcal{P}_f$, $\mathcal{P}_{rf}$}, stop condition $\mathrm{stop}_{align}(\cdot)$
$A^0_{i,n}$ = $\mathcal{M}(C_i, \mathcal{E}_{i,n}; \mathcal{P}_{gen})$
$f^0_{i,n}$ = $\mathcal{M}(A^0_{i,n}, \mathcal{E}_{i,n}; \mathcal{P}_f)$
$t = 0$
$T = 10$
while not $\mathrm{stop}_{align}(f^t_{i,n})$ **and** $t \leq T$ **do**
  $t = t + 1$
  $A^t_{i,n}$ = $\mathcal{M}(C_i, \mathcal{E}_{i,n}, f^{t-1}_{i,n}; \mathcal{P}_{rf})$
  $f^t_{i,n}$ = $\mathcal{M}(A^t_{i,n}, \mathcal{E}_{i,n}; \mathcal{P}_f)$
end while
return $A^t_{i,n}$

4.4. Fact-Checking

The primary objective of fact-checking is to classify the claim into one of five distinct ratings: "True", "Mostly True", "Half True", "Mostly False", or "False". In the context of cross-domain fact-checking, we implement two approaches to serve as the fact-checker in MAGIC to determine the veracity ratings of claims.

Encoder-Based Checker: We utilize a BERT-based model to predict the rating of C_i based on the generated multiple argument $\mathcal{A}_i^{'}$. The XLM-RoBERTa-Base model (Conneau et al., 2020) is employed for multi-classification. The supervised learning setup requires five distinct models, each trained on data from one of the five domains. Thus, we train five XLM-RoBERTa-Base models.

Seq2seq-based checker: In contrast to the encoder-based approach, we use the sequence-to-sequence approach (i.e., seq2seq-based checker) to predict the veracity of the claim in an unsupervised learning manner. Specifically, we evaluate the ability of autoregressive language models to predict ratings. We use an LLM to predict the claim's rating, represented by $y = \mathcal{M}(\mathcal{A}_i; \mathcal{P}_y)$. Here, we prompt $\mathcal{M}$ to produce y from the tokens corresponding to the five possible claim ratings. Thus we do not fine-tune the LLM for fact-checking, and we use a single LLM to predict ratings for claims across all five domains. In MAGIC, we adopt `Vicuna-7b-v1.5` (Zheng et al., 2023) as our seq2seq-based checker.

5. Experiments

5.1. Baseline Models

We use a multi-classification model and a seq2seq model as our baseline models.

RoBERTa: We train the XLM-RoBERTa-Base model without the generated arguments as one of the baselines. That is, the only input is the claim and the associated retrieved evidence. As men-

tioned in Section 4.4, we train five XLM-RoBERTa-Base models for five domains to investigate cross-domain fact-checking. Each model M_k, where $k \in [1, 5]$, is trained exclusively on data within a specific domain. However, since a claim is associated with several retrieved pieces of evidence, and a single piece of evidence $\mathcal{E}_{i,n}$ might exceed the limitation of the model's context window size, we introduce a voting approach for the baseline. In this process, every individual prediction is determined according to evidence $\mathcal{E}_{i,n}$, where $n \in [1, N]$, and N is the total number of C_i's premise articles. The final decision is determined as the rating that receives the majority of votes from these individual predictions. This approach ensures that the most confident prediction is chosen as the final output. To that effect, both training and inference processes follow the voting approach.

In sum, with the retrieved evidence $\mathcal{E}_{i,n}$ of the i-th claim C_i, the fact-checking result based on the voting approach can be formulated as

$$y_i = \text{Mode}(M_k(C_i, \mathcal{E}_{i,1}), M_k(C_i, \mathcal{E}_{i,2}), \ldots, M_k(C_i, \mathcal{E}_{i,N})) \quad (4)$$

where Mode denotes the mode function, used to determine the majority voting outcome.

Zero-shot LLMs: We also benchmark the zero-shot learning performance of the `Vicuna-7b-v1.5` and `GPT-3.5-turbo` LLMs without any fine-tuning. Although LLMs have a larger context window that can handle more evidence, with N potentially reaching values as high as 100, further processing is necessary. Therefore, we also employ the voting approach for zero-shot LLMs.

5.2. Experimental Setup

Dataset: Our dataset comprises 16,177 claims paired with numerous premise articles. The claim and premise article pairs are split into training, validation, and test sets at a ratio of 6:2:2. The validation dataset serves the purpose of evaluating training results, particularly crucial for assessing the performance of encoder-based models. The evaluation results guide us in determining the optimal settings for the model.

In light of our focus on cross-domain fact-checking, we rotate each of the five domains in our dataset to serve as the in-domain data. For each piece of selected in-domain data, we utilize its training set to train both the evidence retriever and the encoder-based checker. The remaining data, based on the predefined ratio, is split into validation and test sets for evaluation. Meanwhile, data from the other four domains are viewed as test sets. If this entire procedure constitutes one round, given our five distinct domains, we conduct five such rounds in total. This experimental setting enables us to evaluate both the in-domain and out-of-domain performance of each method across each round. The goal is to simulate real-world scenarios and assess how well a model trained exclusively on one domain generalizes to unseen domains.

Note that the rating distribution in XClaimCheck is imbalanced, especially after merging "Pants on Fire" with "False". To address this problem, we oversample underrepresented ratings to train all the models used in our experiments.

Hyperparameters: To train the RoBERTa model, we configured the hyperparameters as follows: the learning rate is set at 1×10^{-5}, the batch size at 24, the number of training epochs at 3, and the weight decay at 0.01. Additionally, we shuffled the training dataset with a random seed of 42.

Prompt Format: The prompts employed in this study, denoted as $\mathcal{P}_{gen}$, $\mathcal{P}_f$, and $\mathcal{P}_{refine}$, are detailed in Table 3.

Task	Prompt
Multi-Argument Generation	Output a no more than 50-words argument, utilizing evidence from the reference to assess claim authenticity. Claim: `<Claim>` Claimant: `<Claimant>` Reference: `<Retrieved Evidence>` If unrelated, output "not related."
Feedback	Evaluate if the argument aligns with the facts presented in the reference; if not, provide reasons for the misalignment. Argument: `<Argument>` Reference: `<Retrieved Evidence>`
Self-Refinement	Output a no more than 50-words argument, utilizing evidence from the reference to assess claim authenticity. Incorporate the feedback to ensure alignment. Claim: `<Claim>` Claimant: `<Claimant>` Reference: `<Retrieved Evidence>` Feedback: `<Feedback>` If unrelated, output "not related."

Table 3: Task-oriented prompt templates in MAGIC

5.3. Experimental Results

The performance of each method on cross-domain fact-checking is shown in Table 4. The evaluation metric is the macro-averaged F-score. We report the average macro F-score and the standard deviation across five domains in the "Avg. / std." column. We also report the results for all methods on the in-domain and out-of-domain data. We calculate McNemar's statistical significance test on the baselines and our models. MAGIC (encoder-based) and MAGIC (seq2seq-based) indicate the use of encoder-based and seq2seq-based fact-checkers

Model	Avg. / std.	In-domain	Out-of-domain
RoBERTa	0.2056 ± 0.0228	0.2307	0.1993
Zero-shot Vicuna	0.0667 ± 0.0000	0.0667	0.0667
MAGIC (seq2seq-based)	0.2049 ± 0.0156	0.2012	0.2058
w/o self-refine	0.1842 ± 0.0101	0.1846	0.1841
MAGIC (encoder-based)	**0.2500** ± 0.0175	**0.2661**	**0.2459**
w/o self-refine	0.2391 ± 0.0229	0.2459	0.2374

Table 4: Cross-domain fact-checking results, with the Macro-F1 score serving as the metric

within the framework, respectively. "W/o self-refine" denotes that the self-refinement mechanism is not incorporated in MAGIC.

Overall Fact-Checking Performance: In Table 4, "MAGIC (encoder-based)" significantly outperforms the baseline RoBERTa and "MAGIC (seq2seq-based)" at $p < 0.001$ and $p < 0.001$, respectively. Examination of the impact of the self-refinement mechanism within MAGIC shows that its inclusion clearly benefits both in-domain and out-of-domain fact-checking. Specifically, "MAGIC (encoder-based)" and "MAGIC (seq2seq-based)" outperform the "w/o self-refine" variants (no self-refinement) at $p < 0.001$ and $p < 0.001$ using McNemar's test, respectively.

Multi-Argument Generation: To further examine to what extent multi-argument generation enhances domain generalization, we compare the performance of "Zero-shot Vicuna" and "MAGIC (seq2seq-based) w/o self-refine". The major difference between these two methods lies in their input sources: the former uses the generated multi-argument and the latter relies solely on the retrieved evidence. We find that the performance of "MAGIC (seq2seq-based) w/o self-refine" is higher than that of "Zero-shot Vicuna", indicating that multi-argument generation effectively reconstructs evidence to produce solid arguments. The difference between "MAGIC (encoder-based) w/o self-refine" and RoBERTa, in turn, also lies primarily in the integration of multi-argument generation. The improvement in the performance of "MAGIC (encoder-based) w/o self-refine" over RoBERTa is more pronounced in out-of-domain data than in in-domain data. This significant advancement suggests that multi-argument generation plays a critical role in accurately identifying the veracity of out-of-domain claims.

Domain Dependency: In terms of domain dependency, "MAGIC (encoder-based)" demonstrates a more pronounced disparity between in-domain and out-of-domain performance. In contrast, seq2seq-based approaches such as "MAGIC (seq2seq-based)" exhibit relatively consistent results across in-domain and out-of-domain data. However, the overall performance of "MAGIC (seq2seq-based)" is worse than "MAGIC (encoder-based)". This pilot experiment suggests that the seq2seq-based checker demonstrates low domain dependency, but the zero-shot learning setting with limited parameters (e.g., the 7B LLM) may not be optimal for the task. Fine-tuning the seq2seq-based checker is left as future work.

Table 5 shows a curated set of examples from the constructed XClaimCheck dataset with the corresponding prediction by MAGIC. #1 and #2 denote two different premise articles of the given claim. Below each premise article, the content generated based on that specific article is presented.

6. Discussion

In this section, from the experimental results, we discuss four research questions.

RQ1: How effectively do small LLMs perform across different settings in our method?

We use "small LLMs" to indicate those LLMs that can run on a single machine owned by most organizations, such as `Vicuna-7b-v1.5`. Table 4 shows the results when employing `Vicuna-7b-v1.5` in three different settings. As a fact-checker in zero-shot learning, the small LLM struggles to produce meaningful output. Regardless of the claims and evidence, the model consistently provides simplified responses such as "True", even when probed with detailed prompts.

Nevertheless, when we task the small LLM with generating arguments aligned with the given evidence and determining the ratings (i.e., MAGIC (seq2seq-based)), its ability to predict the veracity of the claim is substantially improved. This implies that compared to making predictions, the small LLM is more proficient at organizing and structuring data. In addition, when provided with a specific structure and context, the small LLM is more adept at producing accurate predictions.

In the absence of fine-tuning, `Vicuna-7b-v1.5`, with 7B parameters, demonstrates commendable capability in argument generation, which substantially benefits subsequent prediction regarding the veracity of claims. This is notably practical for real-world applications, given its cost-effective resource consumption. In the future, there is considerable potential in fine-tuning the model to enhance its performance.

RQ2: How do larger LLMs like GPT-3.5 perform in our setup?

To answer this question, we employ `GPT-3.5-turbo` (Ouyang et al., 2022) with 175B parameters in our experiment. We chose two LLMs with vastly different parameter sizes in order to assess whether a smaller-scale LLM could still achieve commendable performance. The results are shown in Table 6. "ID" and "OOD" denote in-domain and out-of-domain, respectively. In our experiments, we input the results from the evidence retriever to

Topic: Energy	**Claim:** Wind power is the most undependable form of renewable energy.
Short Ruling from Review Article: In terms of energy, dependability can mean different things to different people, and both solar and wind have their strengths and weaknesses. And the actual dependability of each technology is going to rely – to quote the old real estate joke – on three things: location, location and location. The only national ranking we found was done by the U.S. Energy Information Administration, which takes various dependability measures into account. In that ranking, wind is less dependable than hydro, geothermal or biomass but more dependable than solar cells and solar-thermal. O'Neil made an unequivocal and over-arching assertion, but provided no useful support for it. The most convincing evidence we found says he's wrong. We rate his statement False.	
Premise Article #1: Annual Energy Outlook 2021.	**Premise Article #2:** Wind, solar power paired with storage could be cost-effective way to power grid.
Multi-Argument Generation: The claim that wind power is the most undependable form of renewable energy is not supported by the provided reference. The U.S. Energy Information Administration's Annual Energy Outlook 2021 presents a positive outlook for wind power, with projections for increased use and a growing share of the electricity generation mix. The reference does not support the claim that wind power is undependable.	**Multi-Argument Generation:** The claim that wind power is the most undependable form of renewable energy is not supported by the reference provided. The reference discusses a combination of wind power, solar power, and storage as a cost-effective solution to energy demands and states that a well-designed combination of these sources could nearly always exceed electricity demands. The research also found that generating more electricity than needed during average hours to meet demands on high-demand but low-wind power hours would be cheaper than storing excess power for later high demand. The study concludes that a very large electric system can be run almost entirely on renewable energy. Therefore, the claim made by James O'Neil in 2013 that wind power is the most undependable form of renewable energy is not supported by the reference provided.
Alignment Verification: The argument provided does not align with the facts presented in the reference. While the reference does indicate a positive outlook for wind power, with projections for increased use and a growing share of the electricity generation mix, it doesn't directly address the issue of dependability or reliability. Dependability of wind power can refer to its intermittency, variability, and the need for backup power sources when wind generation is low.	**Alignment Verification:** The argument provided aligns with the facts presented in the reference. The argument accurately highlights that a combination of wind power, solar power, and storage can provide a cost-effective solution to energy demands and can potentially exceed electricity demands. This aspect of the argument is supported by the reference, which discusses the potential for a well-designed combination of renewable energy sources to meet energy demands.
Final Argument: Wind power's dependability remains uncertain, as it necessitates backup sources due to its intermittent nature. However, projections from the U.S. Energy Information Administration suggest an increasing share in the electricity mix, indicating potential improvements in reliability over time.	(Since the argument and reference are aligned, there is no need for multiple rounds of argument generation.)
Fact-Checking Result from Encoder-Based Checker: False	

Table 5: Sample demonstrations from XClaimCheck and the corresponding model outputs

Model	Avg. / std.	ID	OOD
MAGIC (encoder-based)	0.2500 $\pm$ 0.0175	0.2661	0.2459
Zero-shot GPT-3.5	0.2606 $\pm$ 0.0319	0.2505	0.2631
w/ encoder-based checker	0.2621 $\pm$ 0.0205	**0.2816**	0.2572
w/ seq2seq-based checker	**0.2623** $\pm$ 0.0250	0.2530	**0.2646**

Table 6: Results using `GPT-3.5-turbo`, with the Macro-F1 score serving as the metric

`GPT-3.5-turbo`, and report the performance on in-domain and out-of-domain data. The third and fourth rows of Table 6 represent settings where `GPT-3.5-turbo` serves as the multi-argument generator paired with an encoder-based checker, and when it acts as both the multi-argument generator and the fact-checker, respectively. In other words, "w/ encoder-based checker" means we input the generated arguments to the encoder-based checker. "w/ seq2seq-based checker" means we use the `GPT-3.5-turbo` to perform fact-checking.

For zero-shot learning, `GPT-3.5-turbo` achieves competitive performance in identifying the veracity of both in-domain and out-of-domain claims. Specifically, when paired with the encoder-based checker, it achieves the highest macro-averaged F-score in in-domain data. By contrast, with the seq2seq-based checker, it achieves the highest macro-averaged F-scores in out-of-domain settings.

In the experiments, we find that the LLM with a large number of parameters indeed exhibits excellent capabilities, whether it is generating arguments or predicting veracity. However, `GPT-3.5-turbo` also exhibits a higher standard deviation in performance across different domains, perhaps suggesting domain knowledge imbalance in handling diverse topics. Nevertheless, MAGIC, even when utilizing a smaller 7B LLM, demonstrates comparable performance, particularly in predicting the veracity of in-domain claims: MAGIC outperforms both "Zero-shot GPT-3.5" and "Zero-shot GPT-3.5 w/ seq2seq-based checker". However, the MAGIC's macro F1 score is slightly lower than that of GPT-3.5 approaches in out-of-domain data.

RQ3: How do various models perform in identifying different claim ratings?

	Rating											
	Pants on Fire		False		Mostly False		Half True		Mostly True		True	
Model	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
RoBERTa	**71.08%**	**64.15%**	**54.38%**	**52.26%**	17.19%	16.02%	17.51%	17.55%	5.80%	6.28%	27.68%	**35.65%**
MAGIC (encoder-based)	52.53%	45.13%	41.34%	35.44%	**28.38%**	**28.55%**	**26.25%**	**17.88%**	28.68%	22.77%	23.37%	28.34%
GPT-3.5 (seq2seq-based)	57.85%	52.31%	40.03%	39.02%	13.89%	14.06%	14.29%	13.10%	**35.33%**	**42.00%**	**27.83%**	30.05%

Table 7: In-domain and out-of-domain accuracy over ratings

	Test domain				
Training domain	Public Policy & Finance	Political Issues	Legal & Regulatory Affairs	Infra. & Services	Global Affairs & Security
Public Policy and Finance	**0.2809**	**0.2809**	0.2584	0.2403	**0.2641**
Political Issues	0.2272	0.2513	**0.2630**	0.2529	0.2414
Legal and Regulatory Affairs	0.2319	0.2471	0.2583	0.2289	0.2484
Infrastructure and Services	0.2503	0.2550	0.2360	**0.2964**	0.2338
Global Affairs and Security	0.2103	0.2348	0.2520	0.2616	0.2435

Table 8: MAGIC (encoder-based) F1 score in cross-domain fact-checking

We select the three top-performing models in our fact-checking task: RoBERTa, MAGIC (encoder-based), and GPT-3.5 (seq2seq-based). In Table 7, we separate the rating "Pants on Fire" from "False", and report the performance of each model for all six ratings. Both RoBERTa and GPT-3.5 achieve better results when identifying the "False" and "True" claims, but they struggle to determine ambiguous ratings such as "Mostly False" and "Half True" claims.

By contrast, MAGIC achieves promising results in assessing claims that fall into these partially true or false categories. The ability to discern partially true or false claims is crucial, especially in real-world contexts. Most individuals can easily verify overtly false or true claims, but evaluating those that are ambiguous demands greater expertise and information access. However, our proposed method still shows room for improvement in discerning "Mostly False" and "Half True" claims. For such cases a more advanced method is needed.

RQ4: How does MAGIC perform in the cross-domain fact-checking task?

Table 8 shows the F1 scores of MAGIC (encoder-based) across various domains. The "Training Domain" and "Test Domain" columns indicate the domain's training set on which the model was trained and the domain's test set evaluated by the model. Note that the best-performing model within each domain is not necessarily trained on that domain's data. For instance, the best results in identifying claims related to "Political issues" are achieved by a model trained on the "Public Policy and Finance" data; indeed, this model consistently excels in several domains. Apart from the model's domain generalization capabilities, the degree of relatedness between the domains' subjects and the complexity of the issues impacts model performance. "Public Policy and Finance", which covers issues closely associated with finance, has significant correlations with other subjects, even when they are categorized within different domains. This may explain why the "Public Policy and Finance" model performs well in other domains as well.

We also observe that the "Infrastructure and Services" domain, consisting of subjects such as technology and energy, often utilizes domain-specific terminology. Because of this, these topics are more self-contained. Therefore, when using a model trained on out-of-domain data to predict a "Infrastructure and Services" claim's rating, the performance is expected to be worse than using a model trained on in-domain data.

7. Conclusion and Future Work

Automatic fact-checking has become a popular research area due to the proliferation of information on the Internet. Verifying the trustworthiness of these claims is important. While some studies investigate various aspects of fact-checking, cross-domain verification remains a challenge. In this work we introduce the task of cross-domain fact-checking and construct XClaimCheck, a dataset consisting of five domains and claims from various topics. To reconstruct the evidence from a vast collection of relevant documents and identify the veracity of claims from unseen domains, we propose MAGIC, a framework for multi-argument generation with self-refinement in fact-checking. Experimental results show that multi-argument generation effectively generates salient arguments from retrieved evidence and that self-refinement enhances the consistency between the generated arguments and the corresponding evidence. However, identifying the veracity of out-of-domain claims remains challenging; this is left as future work. Furthermore, we only use claims from one platform to reflect the cross-domain fact-checking scenario. The diversity of domains is still limited. We plan to expand XClaimCheck in the future.

Acknowledgement

This work was partially supported by the National Science and Technology Council, Taiwan, under grant NSTC 111-2222-E-A49-010-MY2, and by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and the Ministry of Education (MOE), Taiwan.

8. Bibliographical References

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–27.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697.

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 Neural Information Processing Systems*, 33:1877–1901.

Ramón Casillas, Helena Gómez-Adorno, Victor Lomas-Barrie, and Orlando Ramos-Flores. 2022. Automatic fact checking using an interpretable BERT-based architecture on COVID-19 claims. *Applied Sciences*, 12(20):10644.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

L Derczynski, K Bontcheva, M Liakata, R Procter, GWS Hoi, and A Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76. Association for Computational Linguistics.

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.

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

Martin Fajcik, Pavel Smrz, and Lukas Burget. 2019. BUT-FIT at SemEval-2019 Task 7: Determining the rumour stance with pre-trained deep bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1097–1104.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 Task 7: RumourEval 2019: Determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019*, pages 845–854. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of GPT-3. *arXiv preprint arXiv:2209.12356*.

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 Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (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. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. WatClaimCheck: A new dataset for claim entailment and inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. eventAI at SemEval-2019 Task 7: Rumor detection on social media by exploiting content, user credibility and propagation information. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 855–859.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Arya D McCarthy, Hao Zhang, Shankar Kumar, Felix Stahlberg, and Axel H Ng. 2022. Improved long-form spoken language translation with large language models. *arXiv preprint arXiv:2212.09895*.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

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

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. FaVIQ: FAct Verification from Information-seeking Questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5166.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012.

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

Dominik Stammbach. 2021. Evidence selection as a token-level prediction task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 14–20. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. In *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.

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.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting*

of the Association for Computational Linguistics (Volume 2: Short Papers), pages 422–426.

Zhenrui Yue, Huimin Zeng, Ziyi Kou, Lanyu Shang, and Dong Wang. 2022. Contrastive domain adaptation for early misinformation detection: A case study on COVID-19. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2423–2433.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3):e0150989.