Investigating Zero- and Few-shot Generalization in Fact Verification

Liangming Pan[†]
University of California, Santa Barbara
liangmingpan@ucsb.edu

Yunxiang Zhang[†]
University of Michigan, Ann Arbor
yunxiang@umich.edu

Min-Yen Kan
National University of Singapore
kanmy@comp.nus.edu.sg

Abstract

We explore *zero- and few-shot generalization for fact verification (FV)*, which aims to generalize the FV model trained on well-resourced domains (*e.g.*, Wikipedia) to low-resourced domains that lack human annotations. To this end, we first construct a benchmark dataset collection which contains 11 FV datasets representing 6 domains. We conduct an empirical analysis of generalization across these FV datasets, finding that current models generalize poorly. Our analysis reveals that several factors affect generalization, including dataset size, length of evidence, and the type of claims. Finally, we show that two directions of work improve generalization: 1) incorporating domain knowledge via pretraining on specialized domains, and 2) automatically generating training data via claim generation.

1 Introduction

With a rise in deliberate disinformation, *Fact Verification (FV)* has become an important NLP application. FV aims to verify given claims with the evidence retrieved from plain text. Rapid progress has been made by training large neural models (Zhou et al., 2019; Liu et al., 2020; Zhong et al., 2020) on the FEVER dataset (Thorne et al., 2018), containing more than 100K human-crafted (evidence, claim) pairs based on Wikipedia. Fact verification is also needed in other domains, including news, social media, and scientific documents. This has spurred the creation of a large number of FV datasets, such as COVID-Fact (Saakyan et al., 2021), SciFact (Wadden et al., 2020), and Climate-FEVER (Diggelmann et al., 2020).

However, considering that human annotation is time-consuming, costly, and often biased, it is difficult to collect reliable human-labeled data in every domain that demands fact verification. We need to investigate how to build a generalizable fact verification system that adapts to new domains with zero or few samples. Critically, how can we leverage valuable (evidence, claim, label) annotations from rich-resourced domains (*e.g.*, Wikipedia) to aid fact verification in the low-resourced ones (*e.g.*, scholarly documents, and social media)? Although FV datasets have been recently created in different domains, little analysis has shown whether FV models generalize across them and to what extent existing datasets can be leveraged to improve performance in these new domains.

In this paper, we bridge this gap by conducting a comprehensive investigation of *zero- and few-shot generalization in fact verification*. By conducting a holistic study of FV datasets to date, we first carefully select 8 datasets that have artificial or natural claims, human-annotated evidence, and two or three-class labels for our study. We then standardize their data formats as (evidence, claim, label) pairs and create dataset variants with different granularity of evidence, which gives us a total of 11 datasets. We then conduct a thorough empirical study of generalization and transfer across these 11 datasets. We train models on a source dataset, and then evaluate their performance on a target dataset, either without any additional target training examples (*zero-shot setting*) or with a few additional target examples (*few-shot setting*).

We find that RoBERTa-based FV models tend to overfit the particular training set, generalizing poorly to other datasets. Our in-depth analysis shows generalization is related to several key factors, including dataset size, length of evidence, and the claim type. In particular, we find that Wikipedia-based artificial claims (*e.g.*, FEVER) generalize well to natural claims in real-world domains with the growth of dataset size, in contrast to prior work that criticized crowd-sourced claims as having strong annotation bias and being unrepresentative of real-world misinformation (Schuster

[†]Equal Contribution.

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 511–524
November 1–4, 2023.

et al., 2019). Our few-shot generalization experiment further shows that fine-tuning on a small amount of target training data can substantially improve performance.

Armed with the above insights, we explore two ways to improve the generalization of fact verification models. 1) *Domain-specific Pretraining*: initializing the FV model with language models pretrained on specialized domains, and 2) *Data Augmentation*: automatically generating training data for the target domain. Results show that these methods can noticeably improve generalization but still leave unsolved challenges such as inflexibility, high cost, and label consistency.

To the best of our knowledge, this is the first work to perform a thorough investigation of generalization and transfer in fact verification. We open-sourced our dataset collection and codes to support future research towards a universal and robust fact verification system[1].

2 Dataset Curation

In this section, we describe the 11 fact verification datasets included in our study. We first describe the criteria for dataset selection (§ 2.1), and then we introduce the dataset processing (§ 2.2). We show the key characteristics of the datasets in Table 1.

2.1 Dataset Selection

A large number of datasets have recently been introduced to study various tasks for fact-checking, *e.g.*, claim detection, evidence retrieval, fact verification, justification production, etc. Our focus, fact verification, in particular, takes a textual *claim* and a piece of *evidence* as input to predict the *label* for the claim. Let's define these aspects:

• **Claim**: Claims for fact verification are often textual, sentence-level statements, which are categorized into: 1) *real-world natural claims* crawled from dedicated websites, textbooks, forums, etc. 2) *artificial claims* written by crowd-workers.

• **Evidence**: Evidence is the relevant information source for validating the claim. Textual sources, such as news articles, academic papers, and Wikipedia documents, are one of the most commonly used types of evidence. Based on the granularity, we categorize the evidence in existing datasets into: 1) *document-level evidence* such as the Wikipedia page (Thorne et al., 2018), news articles (Hanselowski et al., 2019), and scientific papers (Wadden et al., 2020). 2) *sentence-level evidence* annotated by human experts in the relevant documents to support or refute each claim. 3) *no evidence* is given for each claim; the model needs to retrieve evidence from a large knowledge source.

• **Label**: The label definition for the claim also varies across datasets. The most common definition is the binary label with `supports` and `refutes`, and the three-class label, *i.e.*, `supports/refutes/not enough info`. Some works (Wang, 2017; Augenstein et al., 2019) also employ multi-class labels for more fine-grained degrees of truthfulness (*e.g.* true, mostly-true, mixture, etc), where the number of labels vary greatly, ranging from 4 to 27.

Selection Criteria. We employ the following criteria to select the datasets for our study.

• We consider both natural and artificial claims in various domains.
• We consider the datasets with human-annotated document-level and sentence-level evidence. We exclude datasets without evidence or which provide only non-textual evidence; *i.e.*, tables, knowledge bases, etc.
• We only consider datasets with the binary or the three-class label annotation due to the difficulty of canonicalizing such multi-class labels.

By conducting a holistic study of fact-checking datasets to date, eight different data sources meet our requirements. The full list of candidate datasets we investigate is given in Appendix A.

2.2 Dataset Processing

We then process the selected datasets as follows. 1) We convert each dataset to the unified format of claim-evidence-label triples $(c_i, e_i, l_i)_{i=1}^{N}$. The simplicity of this format allows us to focus on out-of-domain generalization, instead of other orthogonal challenges of fact-checking. 2) We create separate dataset variants by pairing each claim with the evidence in different granularity. This enables us to study the impact of evidence length on generalization. After processing, we obtain the final selection of 11 datasets used. We now briefly introduce the nature of each dataset and its specific processing.

Group I: datasets with *artificial* claims. These are based on Wikipedia articles and are often large in size. However, crowd-sourced claims are often written with minimal edits to reference sentences,

[1] `https://github.com/teacherpeterpan/Fact-Checking-Generalization`

Dataset		Domain	Claim	Evidence	Label	# Claims		Avg. # tokens	
						Train	Test	Claim	Evid.
I	FEVER-sent	Wikipedia	artificial	sent-level	S (52%), R (22%), N (26%)	145,327	19,972	9.4	35.9
	FEVER-para	Wikipedia	artificial	doc-level	S (52%), R (22%), N (26%)	145,327	19,972	9.4	368.7
	VitaminC	Wikipedia	artificial	sent-level	S (50%), R (35%), N (15%)	370,653	63,054	12.6	29.5
	FoolMeTwice	Wikipedia	artificial	sent-level	S (49%), R (51%)	10,419	1,169	15.3	37.0
II	Climate-FEVER-sent	Climate	natural	sent-level	S (25%), R (11%), N (64%)	6,140	1,535	22.8	33.8
	Climate-FEVER-para	Climate	natural	doc-level	S (47%), R (19%), N (34%)	1,103	278	22.9	168.9
	Sci-Fact-sent	Science	natural	sent-level	S (43%), R (22%), N (35%)	868	321	13.8	61.9
	Sci-Fact-para	Science	natural	doc-level	S (43%), R (22%), N (35%)	868	321	13.8	257.3
	PubHealth	Health	natural	sent-level	S (60%), R (36%), N (4%)	8,370	1,050	15.7	137.6
	COVID-Fact	Forum	natural	sent-level	S (32%), R (68%)	3,268	818	12.4	82.5
	FAVIQ	Question	natural	doc-level	S (50%), R (50%)	17,008	4,260	15.2	304.9

Table 1: List of the 11 fact verification datasets for our study and their characteristics.

leading to lexical biases such as the overuse of explicit negation (Schuster et al., 2019).

• **FEVER** (Thorne et al., 2018) asks crowd-workers to mutate sentences from Wikipedia articles to create claims. We use the Wikipedia paragraph associated with each claim as its document-level evidence to construct the *FEVER-para* dataset. We then use the sentence-level gold evidence for the `supports` and `refutes` claims to build the *FEVER-sent* dataset. However, since sentence-level evidence is not available for `NEI` claims, we use the system of Malon (2018) to retrieve the evidence sentences, following Atanasova et al. (2020) and Pan et al. (2021).

• **VitaminC** (Schuster et al., 2021) creates contrastive evidence pairs for each claim, in which evidence pairs are nearly identical in language and content, with the exception that one supports a claim while the other does not.

• **FoolMeTwice** (Eisenschlos et al., 2021) designs a multi-player game that leads to diverse strategies for crafting claims (*e.g.*, temporal inference) based on Wikipedia, resulting in more complex claims with less lexical overlap with the evidence.

Group II: datasets with *natural* claims. These claims are collected from the Internet and then manually verified by professional fact checkers. They represent real-world claims, and originate from diverse domains, such as scholarly documents, news articles, forums, etc. However, due to the difficulty and high cost of manually verifying real-world claims, these datasets are limited in scale.

• **Climate-FEVER** (Diggelmann et al., 2020) consists of 1,535 real-life claims regarding climate-change collected from the Internet. The top five most relevant sentences from Wikipedia are retrieved as the evidence. Humans then annotate each sentence as supporting, refuting, or not enough information to validate the claim. We use the sentence-level annotation as the evidence for each claim to build the *Climate-FEVER-sent*. We construct the document-level evidence for each claim by putting together all of its evidence sentences, which gives us the *Climate-FEVER-para* version.

• **Sci-Fact** (Wadden et al., 2020) consists of 1.4K expert-written scientific claims paired with evidence-containing abstracts annotated with labels and sentence-level rationale. We use the annotated rationale as the sentence-level evidence to build the *Sci-Fact-sent*. We construct the *Sci-Fact-para* version by using the evidence-containing abstract as the document-level evidence for each claim.

• **PubHealth** (Kotonya and Toni, 2020) contains 11.8K claims accompanied by journalist crafted, gold standard judgments to support/refute the claims. The claims are collected from five fact-checking websites, news headlines, and news reviews. We use the judgment texts as evidence to pair with each claim.

• **COVID-Fact** (Saakyan et al., 2021) consists of 4,086 claims concerning the COVID19 pandemic crawled from the *r/COVID19* subreddit. We use their sentence-level evidence annotated by crowd-workers as the evidence.

• **FAVIQ** (Park et al., 2022) contains 26k claims converted from natural ambiguous questions posed by real users. The answer-containing Wikipedia paragraph is provided as the document-level evidence for each claim.

Many of the original datasets do not release their test set. Therefore, we use their original split of train/dev sets as our training and evaluation sets. We also standardize the naming of labels as `supports`, `refutes`, and `NEI`. We visualize the global structure of the datasets with tSNE (van der Maaten and Hinton, 2008) and analyze the domain divergence in Appendix B.

3 Zero/Few-shot Generalization

We now explore the generalization ability of fact verification models across the 11 datasets. We first formulate the task of zero/few-shot generalization.

Task Formulation. Given a claim $\mathcal{C}$ and a piece of evidence $\mathcal{P}$ as inputs, a *fact verification* model $\mathcal{F}$ predicts a label $\mathcal{Y}$ to verify whether $\mathcal{C}$ is supported, refuted, or can not be verified by the information in $\mathcal{P}$. In the *zero-shot generalization* setting, we train models on one source FV dataset, and then evaluate its performance on a target test set, without any additional training data in the target dataset. In the *few-shot generalization* setting, we assume we have a small amount of target training examples.

Fact Verification Model. We use the RoBERTa-large (Liu et al., 2019) as the benchmark model for our study since it has achieved state-of-the-art results in many FV datasets. We concatenate the claim and evidence (`[CLS]` *`claim`* `[SEP]` *`evidence`*) and use it as input for a classification task to predict the label of the claim. We use the `roberta-large` (355M parameters) model provided by the HuggingFace library[2].

3.1 Zero-shot generalization results

Table 2 shows the zero-shot generalization results in macro-averaged F1 for the 3-class fact verification task on all the datasets that have `supports/refutes/NEI` labels, where we partition by dataset group: *Group I, top* (datasets with artificial claims); *Group II, bottom* (datasets with natural claims). In general, the RoBERTa model generalizes poorly in this zero-shot setup. Compared with the in-domain performance (training and testing on the same dataset), the best zero-shot generalization performance shows a large drop of 20.80% on average. This shows that the FV model overfits to the particular dataset and generalizes poorly to unseen datasets. This validates prior work that shows the neural models are brittle when encountering out-of-distribution data. Taking a closer look, we further explore several research questions specific to fact verification behind this general trend.

Do artificial claims and natural claims generalize to each other? The bottom left of Table 2 shows that the model trained on natural claims generalizes badly to datasets with artificial claims, with an average F1 drop of 72% relative to the in-domain performance on the three artificial datasets. In contrast, with natural claims[3], the model generalizes better, with an average F1 drop of 56% (bottom right). This observation supports the argument that artificial claims and natural claims have substantial differences, *e.g.*, Wikipedia vs. real-world domains, high vs. less lexical overlap, and simple vs. diverse reasoning types, as discussed in § 2.2 and related works (Wadden et al., 2020; Saakyan et al., 2021).

However, a surprising and counter-intuitive observation is that the model trained on artificial claims generalizes quite well to natural claims. As shown by the top right section, the average F1 drop narrows to 36.9% when generalizing from artificial to natural claims, markedly better than when generalizing between natural claim datasets (56% average drop). In particular, when trained on *FEVER-sent*, the model achieves the best generalization results on 3 out of 5 datasets with natural claims. However, we will show in the following that the large size of artificial claims contributes a lot to its good generalization performance.

Does generalization improve with more data? To examine whether good generalization on the FEVER and VitaminC datasets comes from their large dataset size, we conducted an experiment controlling for data size. Here, we only take 800 examples for each dataset to train the model. We show the zero-shot generalization results between the five datasets with sentence-level evidence in Table 4. The results on all datasets are shown in Table 10 in Appendix C.

We find that the model trained on natural claim datasets (Group I) can generalize to other natural claims slightly better than the model trained on artificial claim datasets (Group II) in this controlled setting. This confirms that dataset size contributes a lot to generalization ability. Tables 2 and 4 together show that Wikipedia-based artificial claims still generalize well to natural claims in real-world domains with the growth of dataset size, although

[2]https://huggingface.co/roberta-large

[3]For fair comparison, we don't count the dataset pairs with the same data source, *e.g.*, (SciFact-sent, SciFact-para)

Train↓ Test→	FEVER -para	FEVER -sent	VitaminC	C-FEVER -para	C-FEVER -sent	SciFact -para	SciFact -sent	PubHealth
FEVER-para	—	**72.81**	43.87	20.83	40.90	22.09	28.10	9.05
FEVER-sent	**55.57**	—	**62.11**	44.98	**48.70**	**44.98**	**56.15**	21.61
VitaminC	52.04	65.32	—	42.32	44.40	44.14	50.55	**21.97**
C-FEVER-para	17.86	20.04	10.59	—	42.02	29.93	31.62	5.29
C-FEVER-sent	17.87	24.47	20.25	**54.59**	—	25.84	39.39	8.47
SciFact-para	23.96	27.09	28.37	29.85	28.63	—	44.68	6.78
SciFact-sent	16.86	24.50	29.22	20.61	32.50	29.00	—	4.49
PubHealth	35.21	34.41	30.67	34.12	24.18	40.34	42.03	—
SELF	85.58	89.28	86.76	44.61	62.54	52.25	54.27	72.10

Table 2: Macro-F1 of **3-class fact verification** on the evaluation set for all datasets in a zero-shot generalization setup. Rows correspond to the training dataset and columns to the evaluated dataset. The row SELF corresponds to the in-domain performance (training and testing on the same target dataset).

Train↓ Test→	FEVER -para	FEVER -sent	VitaminC	FoolMe Twice	C-FEVER -para	C-FEVER -sent	SciFact -para	SciFact -sent	PubHealth	COVID -Fact	FAVIQ
FEVER-para	—	**94.91**	71.56	72.50	**77.40**	76.04	72.29	75.92	44.75	56.82	55.09
FEVER-sent	**89.08**	—	**79.79**	84.02	74.71	**80.21**	**75.12**	**87.37**	58.25	63.99	61.64
VitaminC	84.62	94.46	—	**84.57**	62.80	54.59	62.37	69.31	55.32	**70.32**	**62.98**
FoolMeTwice	82.58	91.46	78.56	—	71.38	78.24	69.22	84.19	56.81	58.68	59.23
C-FEVER-para	33.37	33.72	52.56	34.15	—	56.66	39.77	40.89	38.61	25.50	33.52
C-FEVER-sent	51.33	62.00	55.91	50.69	75.85	—	66.72	72.08	55.54	43.04	36.53
SciFact-para	33.38	33.57	46.73	34.42	43.35	49.23	—	41.27	38.69	26.64	33.69
SciFact-sent	33.40	33.64	36.91	33.77	42.63	42.46	44.02	—	43.35	26.64	33.51
PubHealth	65.68	64.69	53.57	53.55	53.92	61.78	68.75	71.01	—	40.95	50.89
COVID-Fact	70.94	76.16	37.22	63.02	44.13	51.71	63.60	76.29	**60.06**	—	46.93
FAVIQ	74.57	73.80	59.14	59.67	64.92	60.49	59.08	52.64	40.15	50.25	—

Table 3: F1 of **binary fact verification** on the evaluation set for all datasets in a zero-shot generalization setup. Rows correspond to the training dataset and columns to the evaluated dataset.

Train↓ Test→	FEVER -sent	Vita minC	C-FEVER -sent	SciFact -sent	Pub Health
FEVER-sent	—	22.20	13.66	19.64	24.57
VitaminC	16.93	—	13.78	20.04	**24.98**
C-FEVER-sent	16.63	8.36	—	17.24	2.51
SciFact-sent	27.43	**26.80**	**30.50**	—	13.06
PubHealth	**28.60**	26.69	18.04	**22.22**	—

Table 4: Macro-F1 of 3-class fact verification for all datasets with sentence-level evidence in a zero-shot generalization setup. *The size of training data is controlled to 800 samples for all datasets.*

crowd-sourced claims have been criticized to have strong annotation bias and cannot represent real-life misinformation (Schuster et al., 2019).

Which type of label is more difficult to verify? Table 5 shows the breakdown of the class-wise F1 score. For each dataset, we show the average class-wise F1 when training the model on other datasets (zero-shot) and the class-wise F1 for training on the same dataset (in-domain). The results show that the `refutes` claim has the worst prediction score (in bold) almost for all datasets, in both the zero-shot and the in-domain setting. The in-domain results are in line with the empirical observation that (Jiang et al., 2020) it is often ambiguous to differentiate between `refutes` and `NEI` claims even for trained human annotators. This difficulty still maintains in the zero-shot setting and harms the generalization results.

What is the impact of evidence length? From Table 2, we find that fact verification in a dataset with document-level evidence is more difficult than in the same dataset with sentence-level evidence (an average of 13.29% drop of in-domain F1). This is understandable since document-level evidence requires the model to additionally filter out

Dataset	Zero-shot			In-domain		
	S	R	N	S	R	N
FEVER-para	33.75	**25.85**	34.41	87.05	85.89	**83.81**
FEVER-sent	42.52	**28.24**	44.38	91.32	**89.42**	87.10
VitaminC	50.06	**20.44**	25.96	94.44	89.42	**76.42**
C-FEVER-para	34.14	**24.09**	47.76	68.50	**13.56**	51.76
C-FEVER-sent	28.29	**19.72**	64.00	62.27	**46.40**	78.96
SciFact-para	34.33	**15.35**	51.60	59.29	**23.02**	74.44
SciFact-sent	47.43	**22.23**	55.70	61.87	**18.49**	82.45
PubHealth	20.96	**4.54**	7.79	91.07	82.00	**43.24**

Table 5: Class-wise F1 of 3-class fact verification for the zero-shot generalization setup (left) and the in-domain training setup (right). S: supports; R: refutes; N: NEI.

irrelevant information. Climate-FEVER suffers the largest F1 drop of 31.86%, compared with the slight performance drop on FEVER (4.3%) and SciFact (3.72%). A possible reason is that Climate-FEVER's document-level evidence consists of different (even contradictory) evidence sentences, which requires the model to reason over multiple sentences instead of just selecting the most relevant one.

In terms of generalization, the datasets with sentence-level evidence in general achieve better generalization results than other datasets compared to their doc-level versions. For example, C-FEVER-sent generalizes better than C-FEVER-para on 5 of the 6 datasets excluding themselves. Models trained on sentence-level datasets generalize well to other document-level datasets, but the converse is not true. These results indicate that training the FV model on more fine-grained evidence yields better generalization. This is consistent with the intuition that providing fine-grained evidence eases models' learning in FV, showing the importance of accurate evidence retrieval.

3.2 Zero-shot generalization for binary FV

Many works (Jiang et al., 2020; Saakyan et al., 2021) do not consider `NEI` claims due to their ambiguity. To explore whether our previous observations also hold for the task of *binary fact verification*, we evaluate the generalization results for all 11 datasets using only the `supports` and `refutes` claims for training and evaluation, shown in Table 3. In this setting, artificial claims also generalize well to natural claims in other domains. In 6 of the 7 datasets with natural claims, the best generalization score is from a model trained on artificial claims. This also holds for the evidence length: datasets with sentence-level evidence tend to generalize better than document-level datasets. Finally, compared with the three-class result in Table 2, generalization improves a lot on Climate-FEVER, SciFact, and PubHealth. The reason is that the model struggles in distinguishing between `refutes` and `NEI` claims in these datasets, as reflected by Table 5. Therefore, they benefit a lot from removing the `NEI` label.

Train↓ Test→	C-fever -para	C-fever -sent	SciFact -para	SciFact -sent	Pub Health
FEVER-para	50.04	45.99	59.91	68.18	42.81
FEVER-sent	**55.13**	51.84	**66.12**	**76.39**	40.90
VitaminC	50.41	49.80	58.27	68.59	37.84
SELF-few-shot	22.74	10.75	17.24	33.38	43.62
SELF-full	44.61	**62.54**	52.25	54.27	**72.10**

Table 6: Macro-F1 of three-class fact verification for all datasets in a **few-shot generalization setup**.

3.3 Few-shot generalization results

We now consider the few-shot generalization setting, assuming access to a small number of examples from a target dataset (50 for each class in our experiment). We pre-train a model on a source dataset and then fine-tune it on the target dataset. Our goal is to analyze whether pre-training improves performance compared to training on the target alone.

Table 6 shows the macro-F1 on the evaluation set of all datasets. The rows "SELF-few-shot" and "SELF-full" show the performance of direct training on the 150 samples of the target dataset and the full target training set, respectively (without pre-training on the source dataset). Generally, pre-training on a source FV dataset and fine-tuning to the target outperform "SELF-few-shot" on all 5 datasets and "SELF-full" on 3 out of 5 datasets. This shows that pre-training on a related FV dataset helps to reduce the demand for human-annotated training data in the target domain.

Second, *FEVER-sent* obtains good generalization performance in all evaluation datasets. This strengthens our finding in Section 3.1 that FEVER generalizes well to datasets with natural claims in real-world domains. Last, after finetuning, we see a dramatic improvement in performance compared to Table 2. This highlights that current models over-fit the data they are trained on, and small amounts of data from the target distribution can overcome this generalization gap.

Model	Train↓ Test→	FEVER -para	FEVER -sent	VitaminC	C-FEVER -para	C-FEVER -sent	SciFact -para	SciFact -sent	PubHealth
BERT	FEVER-para	—	64.04	33.82	18.15	29.71	18.53	18.19	3.20
	FEVER-sent	**66.97**	—	**54.75**	35.39	26.49	39.27	39.72	25.95
	VitaminC	54.12	63.28	—	39.57	34.93	40.80	45.51	22.21
BioBERT	FEVER-para	—	67.89	42.41	24.22	38.94	37.69	35.85	8.24
	FEVER-sent	57.18	—	51.95	40.58	39.01	36.83	38.36	37.61
	VitaminC	51.03	60.34	—	**40.60**	**39.72**	43.38	**50.71**	19.44
SciBERT	FEVER-para	—	**68.49**	39.73	20.43	33.84	28.90	35.53	6.37
	FEVER-sent	52.95	—	51.84	35.50	35.68	34.24	39.46	**36.46**
	VitaminC	50.20	58.74	—	37.99	38.79	**43.55**	45.69	20.66

Table 7: Zero-shot generalization performance (macro-F1) when **initialized with different pretraining models**.

4 Improving Generalization

We then investigate two ways to improve the generalization ability of fact verification: 1) incorporating domain knowledge via pretraining on specialized domains, and 2) automatically generating training data via data augmentation.

4.1 Pretraining on Specialized Domains

In-domain knowledge is essential for fact-checking in specialized domains. For example, virology background knowledge is required to verify scientific claims regarding COVID19 (Wadden et al., 2020). When generalizing an FV model from one domain to another, how to endow the model with such in-domain knowledge is a challenging subject worthy of long-term study. Here we explore one simple solution: initializing the model with language models pretrained on specialized domains.

In Table 7, we show the zero-shot generalization performance when initializing the FV model with BioBERT (Lee et al., 2020) (pretrained on biology literature) and SciBERT (Beltagy et al., 2019) (pretrained on scholarly documents). Our goal is to explore whether pretraining on specialized domains helps the generalization. To eliminate the impact of other factors such as the model size, we use the BERT model (Devlin et al., 2019) as the baseline, since BioBERT and SciBERT are both based on the BERT model.

We find that BioBERT and SciBERT both outperform the BERT on the generalization scores in Climate-FEVER, SciFact, and PubHealth, with an average improvement of 21.39% and 12.69% in F1, respectively. However, their performance on Wikipedia-based datasets (FEVER and VitaminC) is relatively worse with BERT (-2.6% and -17.7% for BioBERT and SciBERT, respectively). This confirms the generalization of FV in certain domains (*e.g.*, science) can be improved with the language models pretrained on relevant domains (*e.g.*, scientific papers). We have similar observations for the few-shot generalization setting, shown in Table 11 in Appendix D. Despite the positive results, a suitable pretraining model in certain domains (*e.g.*, tweets) is often unavailable. Moreover, this requires re-training the FV model during domain transfer. Therefore, how to develop a more accessible and less expensive way to incorporate in-domain knowledge required for fact-checking still requires further investigation.

4.2 Data Augmentation

Another direction we explore is improving generalization via data augmentation, which has recently shown promising results in other NLP tasks such as question answering (Yue et al., 2021) and machine translation (Cheng et al., 2020). We first train a *claim generation* model based on the BART (Lewis et al., 2020), using the (evidence, claim, label) triples *in the source domain* as training data. We use the format `[LABEL]` *`label`* `[NER]` *`NERs`* `[EVIDENCE]` *`evidence`* as the input, and use the *`claim`* as the target output for training, where *`NERs`* are the entities appearing in the claim (we add *`NERs`* to guide the model to generate more specific claims). We then apply the trained model to generate claims with different labels in the target domain by separately assigning `supports`, `refutes`, `NEI` as the label prefix of the evidence, and we randomly assign an entity from the evidence as the *`NERs`*[4] to guide the claim generation. We name this method as **BART-gen**.

We train BART-gen on the *FEVER-sent* dataset

[4]Since no ground-truth claim is available for the target domain, the entity cannot come from the claim.

Method↓ Test→	C-fever -para	C-fever -sent	SciFact -para	SciFact -sent	Pub Health
Zero-shot setting					
FEVER-full	44.98	48.70	44.98	**56.15**	21.61
FEVER-control	37.14	45.64	32.42	47.57	20.69
BART-gen	**46.51**	**51.67**	**47.80**	54.63	**69.82**
Few-shot setting					
FEVER-full	**55.13**	51.84	**66.12**	76.39	40.90
FEVER-control	37.17	48.13	62.72	**77.41**	47.03
BART-gen	46.94	**52.80**	50.10	59.00	**70.45**

Table 8: Macro-F1 of three-class fact verification **with data augmentation** (BART-gen) and other baselines. The (top, bottom) shows results for (zero-, few-)shot generalization, respectively.

and generate synthetic training (evidence, claim, label) triples for other datasets. For each piece of evidence in the target domain, we generate six claims with different (`label`, *NERs*) combinations. Table 8 shows the zero- and few-shot generalization results for BART-gen and other baselines: 1) *FEVER-full*: the model is trained on the original FEVER-sent dataset; 2) *FEVER-control*: the model is trained on a random subset of FEVER-sent which has the same amount of data with the generated data; 3) *BART-gen*: the model is trained on the generated data. For the few-shot setting, the model is further fine-tuned with 150 in-domain samples.

For the zero-shot setting results in Table 8, *BART-gen* consistently improves the generalization performance compared with *FEVER-full* (+24.9% in average) and *FEVER-control* (+47.4% in average). The results show that training with generated target data is in general more effective than directly generalizing a model trained on the source data. This is better reflected by comparing *BART-gen* with *FEVER-control* in which the data amount is the same. The improvement is especially noticeable for *PubHealth*, probably because it lacks the `NEI` claims in its original training set. Data augmentation addresses this by generating a sufficiently balanced number of claims for each label.

However, our human evaluation in Appendix E shows that around 30% of generated claims suffer the *label inconsistency* problem, *i.e.*, the BART-gen often generates a fluent claim that does not match our desired label (for example, we want to generate a `refutes` claim, but the generated claim is actually `NEI`). Label inconsistency may introduce conflicting information between the pretraining and fine-tuning stages, which we hypothesize is the cause for the lower level of improvement in fine-tuning the model on the generated data, compared with fine-tuning the model on FEVER. Therefore, although data augmentation is a promising direction to improve generalization, it remains a challenging problem regarding how to generate high-quality claims with consistent labels.

5 Related Work

To overcome the proliferation of misinformation, a great amount of progress has been made in the area of automated fact verification. For modeling, pretraining-based models (Nie et al., 2019; Stammbach and Neumann, 2019; Zhao et al., 2020; Soleimani et al., 2020) have been used for better text representation and have achieved promising performance. Graph-based models (Zhou et al., 2019; Liu et al., 2020; Zhong et al., 2020) are used to facilitate the reasoning over multiple pieces of evidence. However, most existing models rely on large-scale in-domain training data, which is often unrealistic for every domain that demand fact checking. In this paper, we aim to address this by working towards a generalizable fact verification system that can adapt to different domains with zero or few samples in the target domain.

For datasets, various fact-checking datasets representing different real-world domains are proposed, including both naturally occurring (Augenstein et al., 2019; Gupta and Srikumar, 2021; Saakyan et al., 2021; Lu et al., 2023) and human-crafted (Thorne et al., 2018; Sathe et al., 2020; Schuster et al., 2021; Atanasova et al., 2022) fact-checking claims. While these FV datasets focus on different domains, there is still a substantial overlap in the abilities required to verify claims across these datasets. However, little analysis has been done on whether they generalize to one another, and the extent to which existing datasets can be leveraged for improving performance on new ones. Similar studies have been done in other NLP tasks (Talmor and Berant, 2019; Hardalov et al., 2021), while it is less investigated in fact verification. In this paper, we bridge this gap by conducting a comprehensive study of generalization and transfer across existing FV datasets, revealing several key factors for better generalization.

6 Conclusion and Future Work

In this work, we perform a thorough empirical investigation of zero- and few-shot generalization over 11 fact verification datasets. Moreover, we conduct an exhaustive analysis and highlight the most important factors influencing the generalization performance. We further empirically explore two ways to improve generalization in fact verification. We highlight several practical takeaways:

• Overall, the FV model generalizes poorly to unseen datasets compared with in-domain evaluation. However, performance is largely improved by fine-tuning on the target data.

• Artificial claims can also generalize well to natural claims with an increase of dataset size.

• Model trained on sentence-level evidence generalize better than document-level evidence.

• The `refutes` claims are the most difficult to verify among the three labels.

• Domain-specific pretraining and data augmentation consistently improves generalization performance, but they also leave unsolved challenges.

In future work, we plan to experiment with more datasets, including non-English ones. We will also explore the generalization of other sub-tasks in fact-checking, *e.g.*, claim detection, evidence retrieval.

Limitations

In our study, we primarily focused on assessing the generalization capabilities of Transformer-based models, such as RoBERTa. However, we did not extend our evaluation to include zero- and few-shot learning performance on large language models (LLMs) like InstructGPT (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023), due to the high experimental costs. Recently, these LLMs have demonstrated impressive few-shot learning capacities across a variety of natural language processing tasks, including few-shot fact-checking (Pan et al., 2023). However, they are API-based and function as black-box models. This restricts our ability to delve deeper into their behavior, given that we cannot access their model weights or fine-tune them directly. On the contrary, Transformer-based models are open-sourced and replicable, providing a wealth of opportunity for more profound insights into our study and paving the way for future research.

Acknowledgements

This research / project is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: fact extraction and verification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021*.

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

Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. 2020. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177.

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 4684–4696.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3613–3618.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. Advaug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5961–5970.

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 (NAACL-HLT)*, pages 4171–4186.

Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *CoRR*, abs/2012.00614.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan L. Boyd-Graber. 2021. Fool me twice: Entailment from wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 352–365.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1163–1168. The Association for Computational Linguistics.

Ashim Gupta and Vivek Srikumar. 2021. 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 (ACL-IJCNLP)*, pages 675–682.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9011–9028.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Kumar 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.

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.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformation*, 36(4):1234–1240.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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

Xinyuan Lu, Liangming Pan, Qian Liu, Preslav Nakov, and Min-Yen Kan. 2023. SCITAB: A challenging benchmark for compositional reasoning and claim verification on scientific tables. *CoRR*, abs/2305.13186.

Christopher Malon. 2018. Team papelo: Transformer networks at FEVER. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 109–113. Association for Computational Linguistics.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. 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 2553–2566.

Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. CREAK: A dataset for commonsense reasoning over entity knowledge. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 476–483.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6981–7004. Association for Computational Linguistics.

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 (ACL)*, pages 5154–5166.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2173–2178. ACM.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. Covid-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 2116–2129.

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from wikipedia. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6874–6882.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 624–643.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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 3417–3423.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. BERT for evidence retrieval and claim verification. In *Advances in Information Retrieval - 42nd European Conference on IR Research (ECIR)*, volume 12036, pages 359–366.

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

Alon Talmor and Jonathan Berant. 2019. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4911–4921.

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 (NAACL-HLT)*, pages 809–819.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

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.

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 (ACL)*, pages 422–426. Association for Computational Linguistics.

Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. 2021. Contrastive domain adaptation for question answering using limited text corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9575–9593.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul N. Bennett, and Saurabh Tiwary. 2020. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6170–6180.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 892–901.

Dataset	Domain	Claim	Doc-level evidence	Sent-level evidence	NEI claims	Publicly available
FEVER (Thorne et al., 2018)	Wikipedia	artificial	✓	✓	✓	✓
WikiFactCheck (Sathe et al., 2020)	Wikipedia	artificial	✓	✓		✓
HOVER (Jiang et al., 2020)	Wikipedia	artificial	✓	✓		✓
VitaminC (Schuster et al., 2021)	Wikipedia	artificial	✓	✓	✓	✓
Fool Me Twice (Eisenschlos et al., 2021)	Wikipedia	artificial	✓	✓		✓
CREAK (Onoe et al., 2021)	Commonsense	artificial	✓			✓
CreditAccess (Popat et al., 2016)	News	natural	✓	✓		✓
Emergent (Ferreira and Vlachos, 2016)	Emergent	natural	✓	✓	✓	✓
MultiFC (Augenstein et al., 2019)	Multiple	natural			✓	✓
Snopes (Hanselowski et al., 2019)	News	natural	✓	✓	✓	
Climate-FEVER (Diggelmann et al., 2020)	Climate	natural	✓	✓	✓	✓
SciFact (Wadden et al., 2020)	Scientific	natural	✓	✓	✓	✓
PubHealth (Kotonya and Toni, 2020)	Health	natural	✓	✓	✓	✓
COVID-Fact (Saakyan et al., 2021)	Forum	natural	✓	✓		✓
X-Fact (Gupta and Srikumar, 2021)	Multiple	natural	✓		✓	✓
FaVIQ (Park et al., 2022)	Forum	natural	✓	✓		✓

Table 9: A list of candidate fact verification datasets.

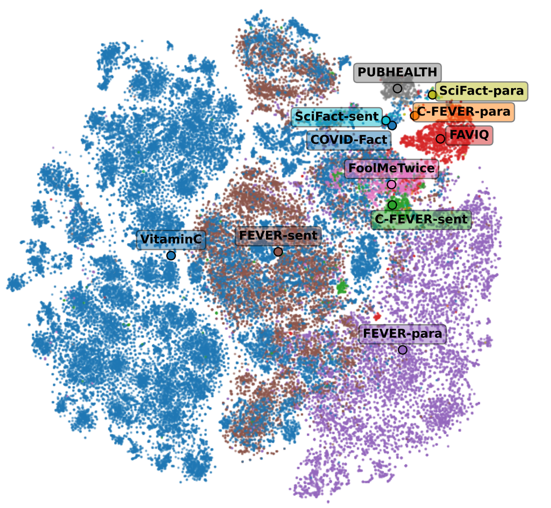


Figure 1: tSNE plot of [CLS] representations of each dataset; highlighted points denote cluster centroids.

A A List of Fact Verification Datasets

In Table 9 we provide a comprehensive list of candidate datasets that we consider for our study, including those are not selected in our benchmark in the end. The candidate list does not include the fact checking datasets without providing evidence for the claim (*e.g.*, FakeNewsNet (Shu et al., 2020)), or focusing on non-textual evidence such as table (*e.g.*, FEVEROUS (Aly et al., 2021) and TabFact (Chen et al., 2020)).

Afterward, we exclude some datasets from the candidate list, mainly because of the lack of clean evidence, the small scale in size, non-English claims, and unavailability. For example, we exclude Emergent (Ferreira and Vlachos, 2016) since it only contains 300 claims. We exclude X-Fact (Gupta and Srikumar, 2021) since it is a multi-lingual dataset that mainly focus on non-English languages. Snopes (Hanselowski et al., 2019) is not included since it is not publicly available. We also exclude CREAK (Onoe et al., 2021), HOVER (Jiang et al., 2020), and MultiFC (Augenstein et al., 2019) since their evidence is either coarse-grained (*e.g.*, the whole Wikipedia page) or noisy (*e.g.*, the original webpage in certain fact checking website).

B Domain Divergence Analysis

Following Hardalov et al. (2021), we plot the 11 datasets in a latent vector space to visualize the global structure of the datasets. We proportionally sample 82K (10%) examples, and we pass them through a RoBERTa-large (Liu et al., 2019) model without any training. The input has the following form: `[CLS] claim [SEP] evidence`. Next, we take the `[CLS]` token representations, and we plot them in Figure 1 using tSNE (van der Maaten and Hinton, 2008). We can see that datasets with *natural claims* are grouped top-right, clearly

Train↓ Test→	FEVER -para	FEVER -sent	VitaminC	C-FEVER -para	C-FEVER -sent	SciFact -para	SciFact -sent	PubHealth
FEVER-para	—	**31.98**	27.58	32.83	25.22	**25.36**	30.72	18.82
FEVER-sent	16.68	—	22.20	21.78	13.66	20.01	19.64	24.57
VitaminC	16.70	16.93	—	22.01	13.78	20.04	20.04	**24.98**
C-FEVER-para	17.15	18.00	27.51	—	**31.44**	17.24	18.19	5.04
C-FEVER-sent	16.63	16.63	8.36	17.51	—	17.24	17.24	2.51
SciFact-para	**30.66**	30.66	**28.30**	**33.46**	31.36	—	**42.40**	12.15
SciFact-sent	28.37	27.43	26.80	27.93	30.50	24.51	—	13.06
PUBHEALTH	26.77	28.60	26.69	28.25	18.04	22.80	22.22	—
SELF	32.35	16.68	29.53	36.12	26.14	53.33	50.72	52.05

Table 10: Macro-F1 of **3-class fact verification** on the evaluation set for all datasets in a zero-shot generalization setup. The size of training data is controlled to 800 samples for all datasets. Rows correspond to the training dataset and columns to the evaluated dataset. The row SELF corresponds to the in-domain performance (training and testing on the same target dataset).

separated from those with *artificial claims*. The clusters of real-world domain datasets do not overlap, which highlights the rich diversity of our selected datasets. We also notice that datasets with *sentence-level* evidence have little overlap with their *paragraph-level* counterparts (e.g., Climate-FEVER-sentence v.s. Climate-FEVER-paragraph). To sum up, Figure 1 confirms that there exists divergence between different domains and datasets.

C Full Results of Controlled Size Generalization

Table 10 shows the full results of the controlled experiment in Section 3.1 where we only take 800 examples for each dataset to train the model. We find that the model trained on artificial claim datasets generalize slightly worse to natural claims compared with the model trained on artificial claim datasets in the controlled size setting. Compared with the good generalization results from artificial claims to natural claims in Table 2, it shows that the size of the source dataset contributes a lot to generalization ability of fact verification.

D Full Results of Few-shot Generalization

In Table 11, we show the few-shot generalization performance of FV model pretrained on specialized domains. After finetuning, we observe dramatic improvement in performance comparing to Table 7 (+14.31% for BERT, +11.84% for BioBERT, +15.29% for SciBERT). Under few-shot setting, we find that BioBERT and SciBERT still outperform the BERT on the generalization scores in all datasets except Climate-FEVER-sentence.

E Human Evaluation of Generated Claims

We conduct the human evaluation on the claims generated by *BART-gen* on four datasets: Climate-FEVER-sentence, Sci-Fact-sentence, PubHealth, and COVID-Fact. We randomly sample 90 generated claims for each dataset with a balanced desired label distribution. To be specific, 30/30/30 of their *desired labels*, *i.e.*, the type of claim we expect the model to generate (by appending the corresponding label prefix) are `supports`/`refutes`/`NEI`. We ask two expert human annotators to annotate the *actual label* for each claim, *i.e.*, whether the generated claim is supported, refuted, or cannot be verified by the evidence. If the generated claim is an incomplete or unreadable sentence, we label it as `Unclassified`.

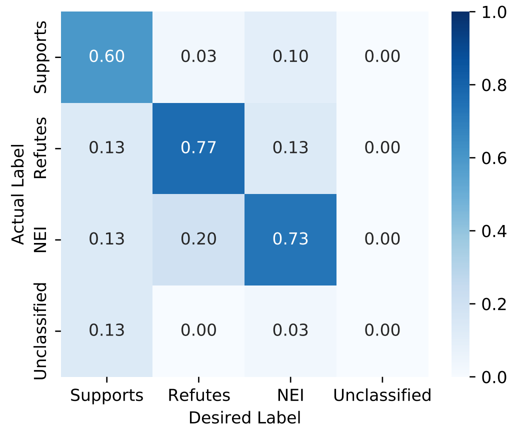
Figure 2 shows the confusion matrix for the desired labels and the actual labels. We find that in all four datasets, around 30% of the generated claims suffer from the *label inconsistency* problem, *i.e.*, the actual label of the claim is not the desired label. Specially, the confusion between the `refutes` and `NEI` claim is the major type of error, showing that `refutes` and `NEI` claims are the hardest for the model to generate.

We also observe that around 5% of the generated claims are incomplete or unreadable. Moreover, most generated claims are short and simple (e.g., "*Gilbert Rothschild was a person*"), which do not require complex reasoning to verify. It is therefore worthy to investigate how to obtain high-quality claims in data augmentation for better generalization in the future study.

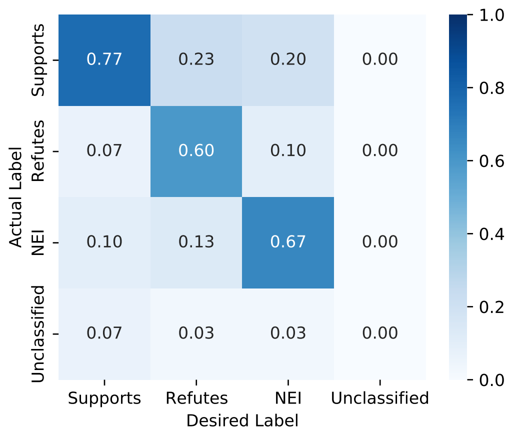
Model	Train↓ Test→	C-FEVER -para	C-FEVER -sent	SciFact -para	SciFact -sent	PubHealth
BERT	FEVER-para	41.24	42.68	42.74	43.42	15.57
	FEVER-sent	43.84	**45.09**	50.23	55.65	33.45
	VitaminC	45.94	43.38	57.25	58.12	33.69
BioBERT	FEVER-para	46.22	43.06	**61.45**	59.19	33.47
	FEVER-sent	47.64	43.43	48.59	53.17	39.44
	VitaminC	44.16	40.48	54.52	**61.62**	32.29
SciBERT	FEVER-para	**47.92**	40.46	53.12	56.83	34.39
	FEVER-sent	43.92	40.94	56.53	60.49	**42.14**
	VitaminC	46.23	40.16	60.95	61.04	37.37

Table 11: Few-shot generalization performance (macro-F1) when **initialized with different pretraining models**.

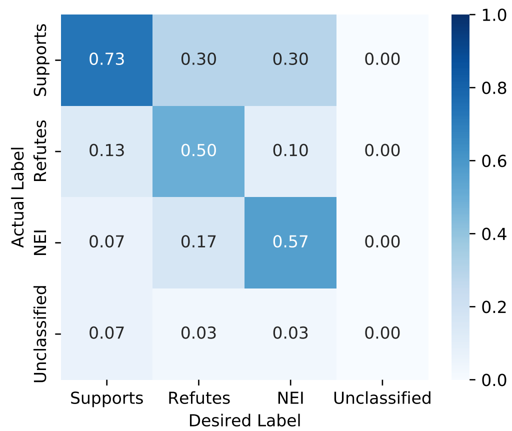
(a) Climate-FEVER-sentence



(b) SciFact-sentence



(c) PubHealth



(d) COVID-Fact

Figure 2: Confusion matrices (normalized over columns) of generated claims on four datasets. The desired label is the input label for our claim generation model (BART-gen) and the actual label is the human-annotated label. "Unclassified" means that the generated claim is incomplete or unreadable.