

# MultiVerS

Improving scientific claim verification with weak supervision and full-document context

**CS431.O12.KHCL**

*Instructor: Nguyen Duy Khanh*

**Group 13**

Le Gia Khang    Nguyen Hoang Tan    Le Duy Khang

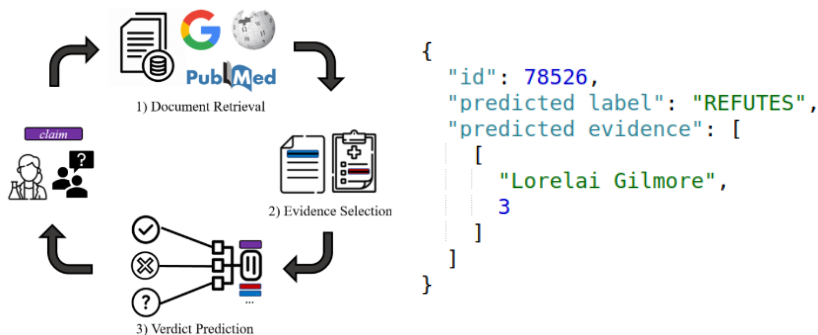
December 30, 2023

## Definition of scientific claim verification from the SCIFACT task:

Given a claim  $c$  and a *candidate abstract*  $a$ .

Label  $y(c, a) \in \{SUPPORTS, REFUTES, NEI\}$ .

Identify rationales  $R(c, a) = \{r_1(c, a), \dots, r_n(c, a)\}$ .



**Claim:**

Advil (ibuprofen) worsens  
COVID-19 symptoms

**Evidence abstract:****Covid-19 and avoiding  
Ibuprofen.**

...

Increased risk of COVID-19  
infection was feared with  
ibuprofen use

...

At this time, there are no  
findings discouraging the use  
of ibuprofen

**Claim:**

Advil (ibuprofen) worsens  
COVID-19 symptoms

**Label: Refuted**

**Evidence abstract:****Covid-19 and avoiding  
Ibuprofen.**

...

Increased risk of COVID-19  
infection was feared with  
ibuprofen use

...

At this time, there are no  
findings discouraging the use  
of ibuprofen

## Claim:

Advil (ibuprofen) worsens  
COVID-19 symptoms

**Label: Refuted**

## Task Outputs

- 1 Fact-checking label

## Evidence abstract:

### Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19  
infection was feared with  
ibuprofen use

...

At this time, there are no  
findings discouraging the use  
of ibuprofen

## Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

**Label: Refuted**

## Task Outputs

- ① Fact-checking label
- ② Rationales justifying the label

## Evidence abstract:

### Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19 infection was feared with ibuprofen use

...

At this time, there are no findings discouraging the use of ibuprofen

**Rationale**

## Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

**Label: Refuted**

## Task Outputs

- ① Fact-checking label
- ② Rationales justifying the label

## Evidence abstract:

### Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19 infection was feared with ibuprofen use

Context required

...

At this time, there are no findings discouraging the use of ibuprofen

Rationale

## Claim:

Advil (ibuprofen) worsens  
COVID-19 symptoms

## Evidence abstract:

### Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19  
infection was feared with  
ibuprofen use

...

At this time, there are no  
findings discouraging the use  
of ibuprofen



# Prior work: Extract-then-label

## Claim:

Advil (ibuprofen) worsens  
COVID-19 symptoms

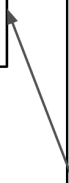
At this time, there are no  
findings discouraging the use  
of ibuprofen

## Evidence abstract:

### Covid-19 and avoiding Ibuprofen.

...  
Increased risk of COVID-19  
infection was feared with  
ibuprofen use

...  
At this time, there are no  
findings discouraging the use  
of ibuprofen



# Prior work: Extract-then-label

## Claim:

Advil (ibuprofen) worsens COVID-19 symptoms

At this time, there are no findings discouraging the use of ibuprofen



**Label: Refuted**

## Evidence abstract:

### Covid-19 and avoiding Ibuprofen.

...  
Increased risk of COVID-19 infection was feared with ibuprofen use

...  
At this time, there are no findings discouraging the use of ibuprofen

# Prior work: Extract-then-label

## Claim:

Advil (ibuprofen) worsens  
COVID-19 symptoms

At this time, there are no  
findings discouraging the use  
of ibuprofen



**Label: Refuted**

## Evidence abstract:

### Covid-19 and avoiding Ibuprofen.

...

Increased risk of COVID-19  
infection was feared with  
ibuprofen use

...

At this time, there are no  
findings discouraging the use  
of ibuprofen

## Drawbacks of extract-then-label:

- ① Rationales may lack context
- ② Requires rationale supervision during training

Given a claim  $c$  and candidate abstract  $a$

**These models make predictions in 2 steps:**

Predict rationales  $\hat{R}(c, a) = \{\hat{r}_1(c, a), \dots \hat{r}_n(c, a)\}$

Then, make a label prediction  $\hat{y}(c, f_R(\hat{R}(c, a)))$

## A multitask system for full-context scientific claim verification

- Predict  $\hat{y}(c, a)$  directly based on an encoding of the entire claim and abstract.
- Enforce consistency of  $\hat{R}(c, a)$  with  $\hat{y}(c; a)$  during decoding.

## Long document encoding:

A claim  $c$  and candidate abstract  $a$  consisting of title  $t$  and sentences  $s_1, \dots, s_n$ .

The  $\langle /s \rangle$  token following each sentence  $s_i$  is notated as  $\langle /s \rangle_i$ .

$$\langle s \rangle \ c \ \langle /s \rangle \ t \ \langle /s \rangle \ s_1 \ \langle /s \rangle_1 \ \dots s_n \ \langle /s \rangle_n$$

Global attention is assigned to  $\langle s \rangle$  token, all tokens in  $c$  and all  $\langle /s \rangle$  tokens.

## Claim

Advil (ibuprofen)  
worsens COVID-19  
symptoms

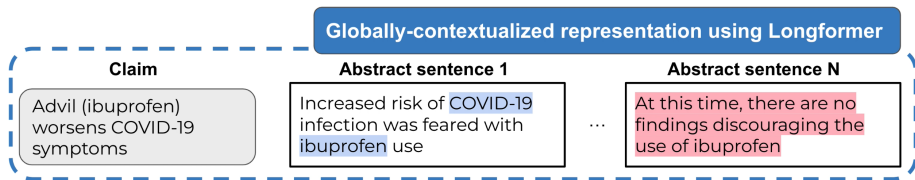
## Abstract sentence 1

Increased risk of COVID-19  
infection was feared with  
ibuprofen use

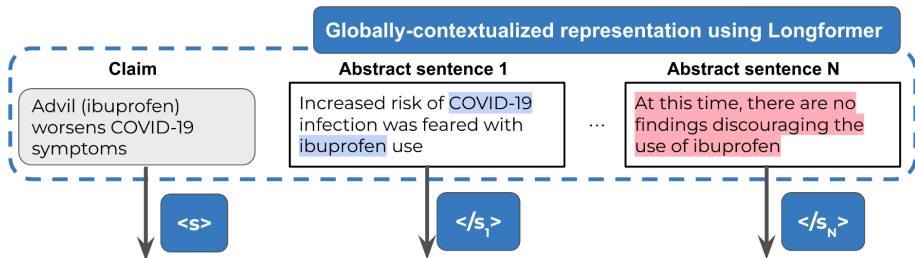
...

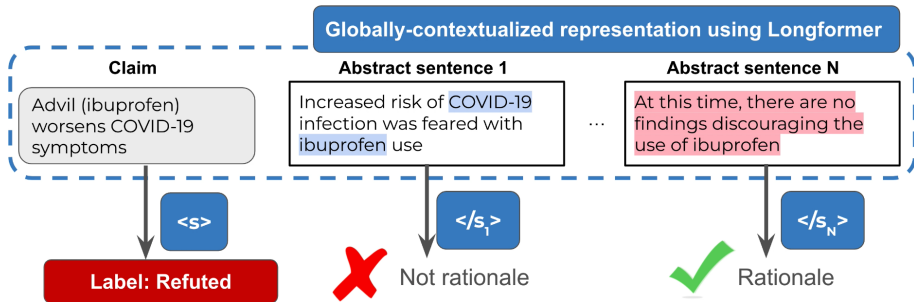
## Abstract sentence N

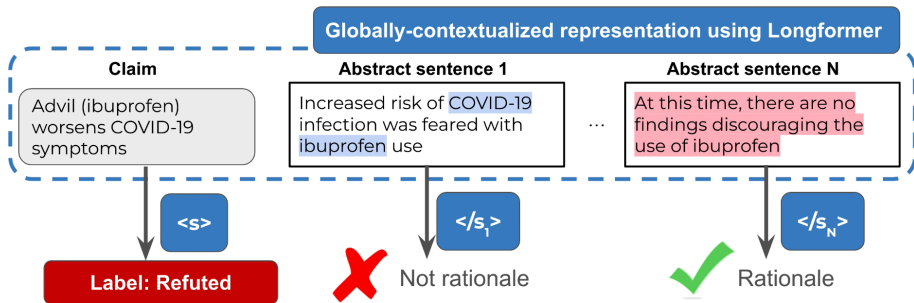
At this time, there are no  
findings discouraging the  
use of ibuprofen











$$\mathcal{L} = \mathcal{L}_{label} + \lambda_{rationale} \mathcal{L}_{rationale}$$

## Benefits of multitask approach:

- ① Incorporates all relevant context
- ② Can train on instances with no rationale annotations

# Experiments

Dataset	Domain	Claim source	Open	Has NEI	Claim complexity	Negation method	Train claims	Eval claims	> 512 tokens
HealthVer	COVID	TREC-COVID	✗	✓	Complex	Natural	1,622	230	14.9%
COVIDFact	COVID	Reddit	✗	✗	Complex	Automatic	903	313	12.4%
SCIFACT	Biomed	Citations	✓	✓	Atomic	Human	1,109	300	27.4%
FEVER	Wiki	Wikipedia	-	✓	Atomic	Human	130,644	-	33.2%
PUBMEDQA	Biomed	Paper titles	-	✓	Complex	Automatic	58,370	-	12.1%
EVIDENCEINFERENCE	Biomed	ICO prompts	-	✓	Atomic	Automatic	7,395	-	42.7%

Table: Summary of datasets used in experiments

# Experiments

## Target datasets:

- HealthVer
- COVID-Fact
- SciFact

# Experiments

## Target datasets:

- HealthVer
- COVID-Fact
- SciFact

Roughly 1000 claims / dataset.

Expert annotations are expensive

# Experiments

## Target datasets:

- HealthVer
- COVID-Fact
- SciFact

Roughly 1000 claims / dataset.

Expert annotations are expensive

## Training procedure:

- **Stage 1:** Train on a combination of *labeled out of domain data* and *weakly-labeled in-domain data*.
- **Stage 2:** Continue training on data from each target dataset.

# Experiments

## Target datasets:

- HealthVer
- COVID-Fact
- SciFact

Roughly 1000 claims / dataset.

Expert annotations are expensive

## Traning procedure:

- **Stage 1:** Train on a combination of *labeled out of domain* data *weakly-labeled in-domain* data.
- **Stage 2:** Continue training on data from each target dataset.

## Domain adaptation settings:

- **Zero-shot:** Stage 1 training only.
- **Few-shot:** 45 instances from target datasets.
- **Full-supervised:** All target data.



**Supervised** out-of-domain  
data (FEVER)

LeBron James was born in  
Ohio

Label: Supported

LeBron James is an American  
basketball player. He was born  
in Akron, Ohio.

# Data: Stage 1

## Supervised out-of-domain data (FEVER)

LeBron James was born in Ohio

Label: Supported

LeBron James is an American basketball player. He was born in Akron, Ohio.

## Weakly-supervised in-domain data

**Diabetes increases risk of depression**

Abstract

...  
...  
...

Label: Supported

# Data: Stage 1

## Supervised out-of-domain data (FEVER)

LeBron James was born in Ohio

Label: Supported

LeBron James is an American basketball player. He was born in Akron, Ohio.

## Weakly-supervised in-domain data

**Diabetes increases risk of depression**

} Claim: Paper title

Abstract

...  
...  
...

Label: Supported

# Data: Stage 1

## Supervised out-of-domain data (FEVER)

LeBron James was born in Ohio

Label: Supported

LeBron James is an American basketball player. He was born in Akron, Ohio.

## Weakly-supervised in-domain data

**Diabetes increases risk of depression**

Abstract

...  
...  
...

Label: Supported

Claim: Paper title

**Rationales** likely to appear in abstract, but are not annotated

# Data: Stage 1

## Supervised out-of-domain data (FEVER)

LeBron James was born in Ohio

Label: Supported

LeBron James is an American basketball player. He was born in Akron, Ohio.

## Weakly-supervised in-domain data

**Diabetes increases risk of depression**

Abstract

...  
...  
...

Label: Supported

Claim: Paper title

**Rationales** likely to appear in abstract, but are not annotated

MultiVerS can train on these examples, even though no rationale annotations are provided

## Abstract-level evaluation:

- Identifying abstracts that SUPPORT or REFUTE each claim.
- Predicting the correct label  $y(c, a)$  is sufficient

## Sentence-level evaluation:

- It combines the accuracy of abstract-level label prediction with the precision of rationale identification.

Setting	Model	HealthVer						COVIDFact						SciFact					
		Abstract			Sentence			Abstract			Sentence			Abstract			Sentence		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Zero	PARAGRAPHJOINT	72.3	14.4	24.0	22.9	2.7	4.9	51.3	37.9	43.6	31.5	16.0	21.3	52.9	32.4	40.2	36.4	14.9	21.1
	MULTIVERs	60.6	20.5	<b>30.7</b>	25.0	4.6	<b>7.8</b>	48.8	45.7	<b>47.2</b>	32.7	18.5	<b>23.6</b>	49.0	44.6	<b>46.7</b>	39.0	21.6	<b>27.8</b>
Few	PARAGRAPHJOINT	62.7	41.6	50.0	46.0	29.3	<b>35.8</b>	73.3	60.6	66.3	44.3	30.6	36.2	44.4	51.4	47.6	33.0	35.1	34.0
	MULTIVERs	63.6	47.9	<b>54.7</b>	41.9	31.0	35.7	71.3	68.1	<b>69.7</b>	39.5	35.4	<b>37.4</b>	76.4	54.1	<b>63.3</b>	51.7	40.3	<b>45.3</b>
Full	VERT5ERINI	71.3	74.0	72.6	65.6	61.2	63.3	76.6	52.7	62.4	44.8	27.2	33.9	64.0	73.0	68.2	60.6	66.5	63.4
	PARAGRAPHJOINT	75.0	68.3	71.5	69.9	60.6	64.9	71.5	68.1	69.8	41.4	40.3	40.8	75.8	63.5	69.1	68.9	54.6	60.9
	MULTIVERs	78.9	76.3	<b>77.6</b>	71.4	67.0	<b>69.1</b>	77.3	77.3	<b>77.3</b>	41.5	46.1	<b>43.7</b>	73.8	71.2	<b>72.5</b>	67.4	67.0	<b>67.2</b>

Table 2: Performance of MultiVerS and baselines.

# Our Reproduction Performance

**covidfact\_covidfact.jsonl:**

```
{  
  "abstract_label_only_precision": 0.7728706624605678,  
  "abstract_label_only_recall": 0.7728706624605678,  
  "abstract_label_only_f1": 0.7728706624605678,  
  "sentence_label_precision": 0.41694915254237286,  
  "sentence_label_recall": 0.458955223880597,  
  "sentence_label_f1": 0.43694493783303723  
}
```



# Our Reproduction Performance

**covidfact\_fever.jsonl:**

```
{  
  "abstract_label_only_precision": 0.9583333333333334,  
  "abstract_label_only_recall": 0.14511041009463724,  
  "abstract_label_only_f1": 0.25205479452054796,  
  "sentence_label_precision": 0.631578947368421,  
  "sentence_label_recall": 0.04477611940298507,  
  "sentence_label_f1": 0.08362369337979093  
}
```

# Our Reproduction Performance

**covidfact\_fever\_sci.jsonl:**

```
{  
  "abstract_label_only_precision": 0.4882154882154882,  
  "abstract_label_only_recall": 0.45741324921135645,  
  "abstract_label_only_f1": 0.4723127035830619,  
  "sentence_label_precision": 0.5444444444444444,  
  "sentence_label_recall": 0.0914179104477612,  
  "sentence_label_f1": 0.15654952076677317  
}
```

# Our Reproduction Performance

**healthver\_fever.jsonl:**

```
{  
  "abstract_label_only_precision": 0.8,  
  "abstract_label_only_recall": 0.00667779632721202,  
  "abstract_label_only_f1": 0.013245033112582783,  
  "sentence_label_precision": 1.0,  
  "sentence_label_recall": 0.0027223230490018148,  
  "sentence_label_f1": 0.005429864253393665  
}
```

# Our Reproduction Performance

**healthver\_fever\_sci.jsonl:**

```
{  
  "abstract_label_only_precision": 0.6059113300492611,  
  "abstract_label_only_recall": 0.20534223706176963,  
  "abstract_label_only_f1": 0.30673316708229426,  
  "sentence_label_precision": 0.40625,  
  "sentence_label_recall": 0.011796733212341199,  
  "sentence_label_f1": 0.022927689594356263  
}
```

# Our Reproduction Performance

**healthver\_healthver.jsonl:**

```
{  
  "abstract_label_only_precision": 0.7892918825561313,  
  "abstract_label_only_recall": 0.7629382303839732,  
  "abstract_label_only_f1": 0.7758913412563666,  
  "sentence_label_precision": 0.7194525904203324,  
  "sentence_label_recall": 0.6678765880217786,  
  "sentence_label_f1": 0.6927058823529412  
}
```

# Our Reproduction Performance

**scifact\_fever.jsonl:**

```
{  
  "abstract_label_only_precision": 0.8378378378378378,  
  "abstract_label_only_recall": 0.13963963963963963,  
  "abstract_label_only_f1": 0.23938223938223935,  
  "sentence_label_precision": 0.7307692307692307,  
  "sentence_label_recall": 0.051351351351351354,  
  "sentence_label_f1": 0.09595959595959597  
}
```

# Our Reproduction Performance

**scifact\_fever\_sci.jsonl:**

```
{  
  "abstract_label_only_precision": 0.4900990099009901,  
  "abstract_label_only_recall": 0.44594594594594594,  
  "abstract_label_only_f1": 0.4669811320754717,  
  "sentence_label_precision": 0.7017543859649122,  
  "sentence_label_recall": 0.10810810810810811,  
  "sentence_label_f1": 0.1873536299765808  
}
```

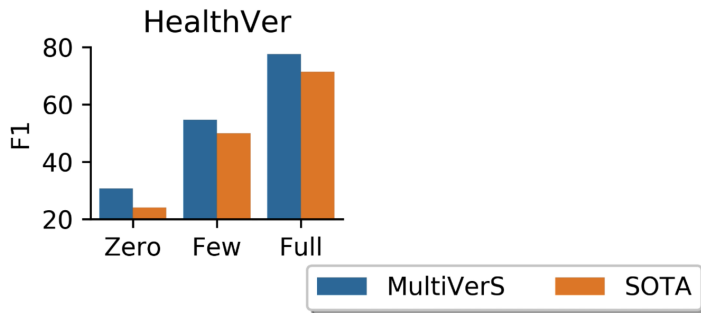
# Our Reproduction Performance

**scifact\_scifact.jsonl:**

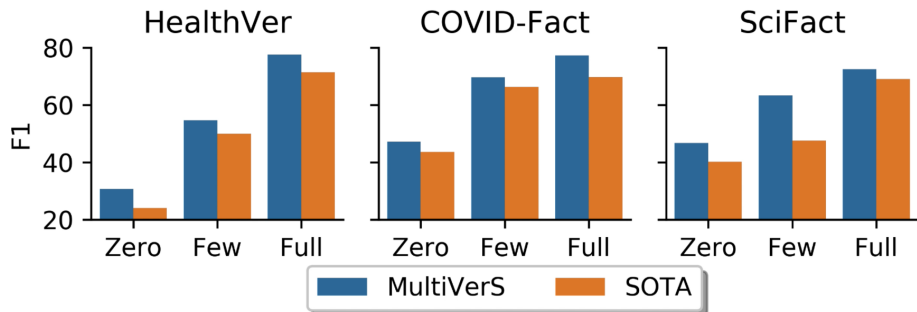
```
{  
  "abstract_label_only_precision": 0.7383177570093458,  
  "abstract_label_only_recall": 0.7117117117117117,  
  "abstract_label_only_f1": 0.7247706422018348,  
  "sentence_label_precision": 0.6739130434782609,  
  "sentence_label_recall": 0.6702702702702703,  
  "sentence_label_f1": 0.6720867208672086  
}
```



# Results

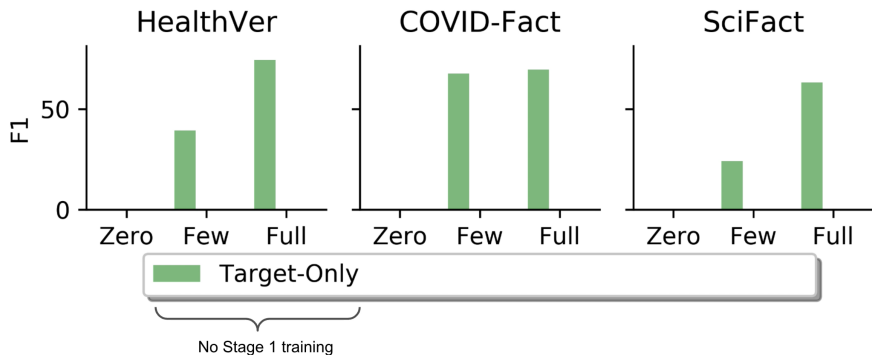


# Results

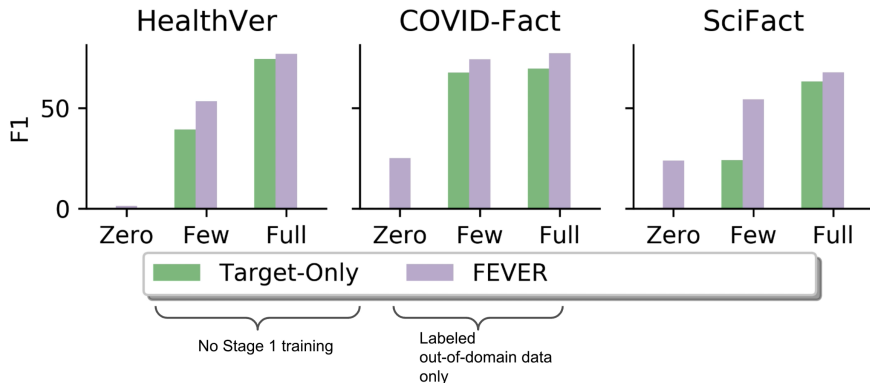


MultiVerS outperforms SOTA on all datasets

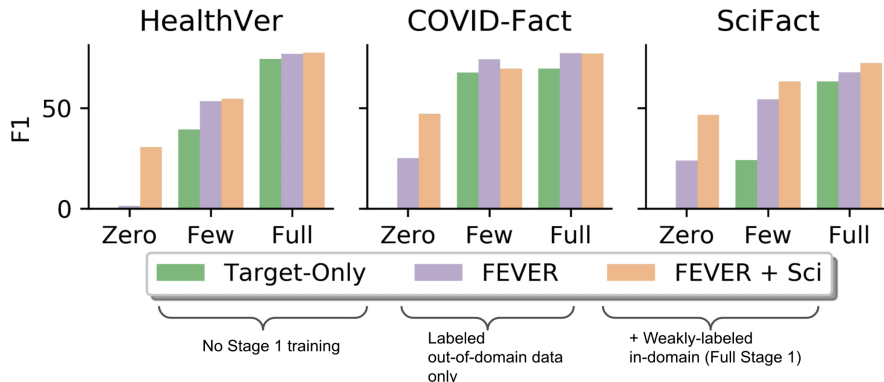
# Ablations: Training strategy



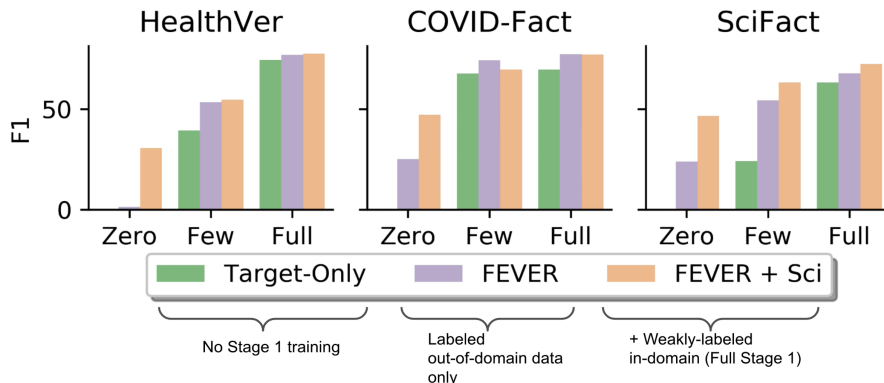
# Ablations: Training strategy



# Ablations: Training strategy



# Ablations: Training strategy



Pretraining with weakly-supervised in-domain data improves few / zero shot performance.

# Reference



MULTIVERS: Improving scientific claim verification with weak supervision and full-document context



Scientific Fact-Checking: A Survey of Resources and Approaches  
Juraj Vladika and Florian Matthes



Longformer: The Long-Document Transformer  
Iz Beltagy, Matthew E. Peters and Arman Cohan



[Code and model checkpoints for the MultiVerS model](#)  
[dwadden/multivers](#)

# Thanks for listening!

Q&A section