2024 AAAI

W17: FACTIFY 3.0 – Workshop Series on Multimodal Fact-Checking and Hate Speech Detection

Team Trifecta at Factify5WQA: Setting the Standard in Fact Verification with Fine-Tuning

Shang-Hsuan Chiang, Ming-Chih Lo, Lin-Wei Chao and Wen-Chih Peng

Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan




Introduction

Problem

Input:

Claim && Evidence && Question

Output Label:

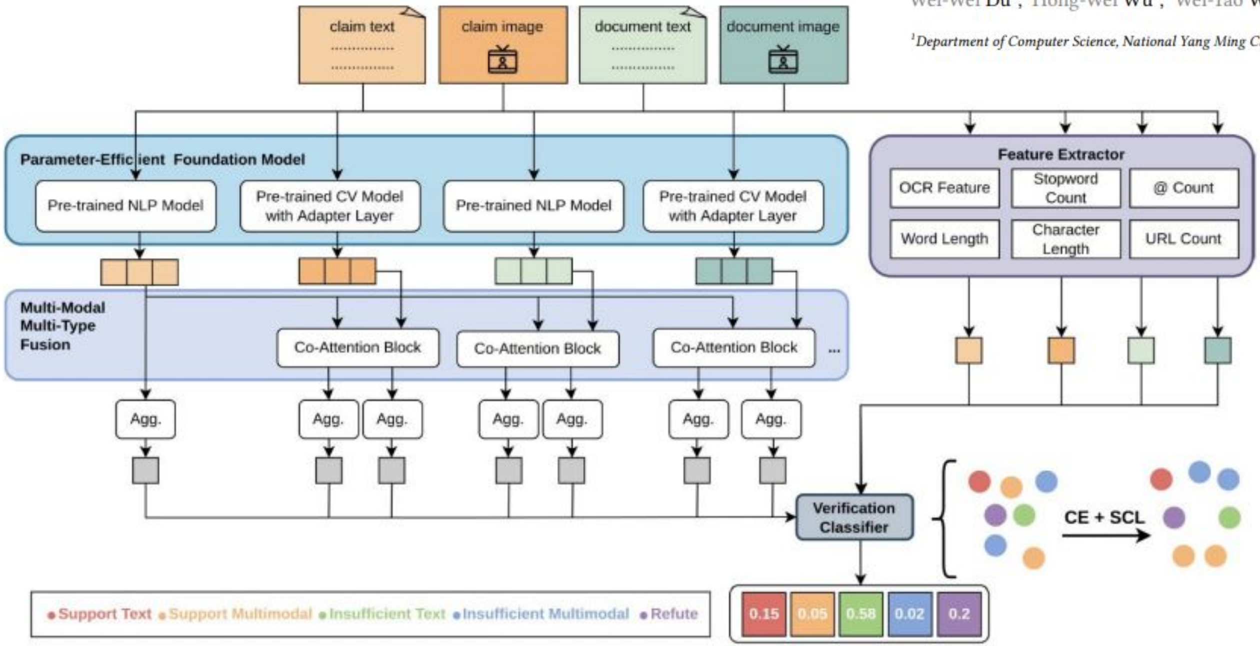
Support || Refute || Neutral

Previous Solution: Pre-CoFactv2

Team Triple-Check at Factify 2: Parameter-Efficient Large Foundation Models with Feature Representations for Multi-Modal Fact Verification

Wei-Wei Du[1], Hong-Wei Wu[1], Wei-Yao Wang[1] and Wen-Chih Peng[1]

[1]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan



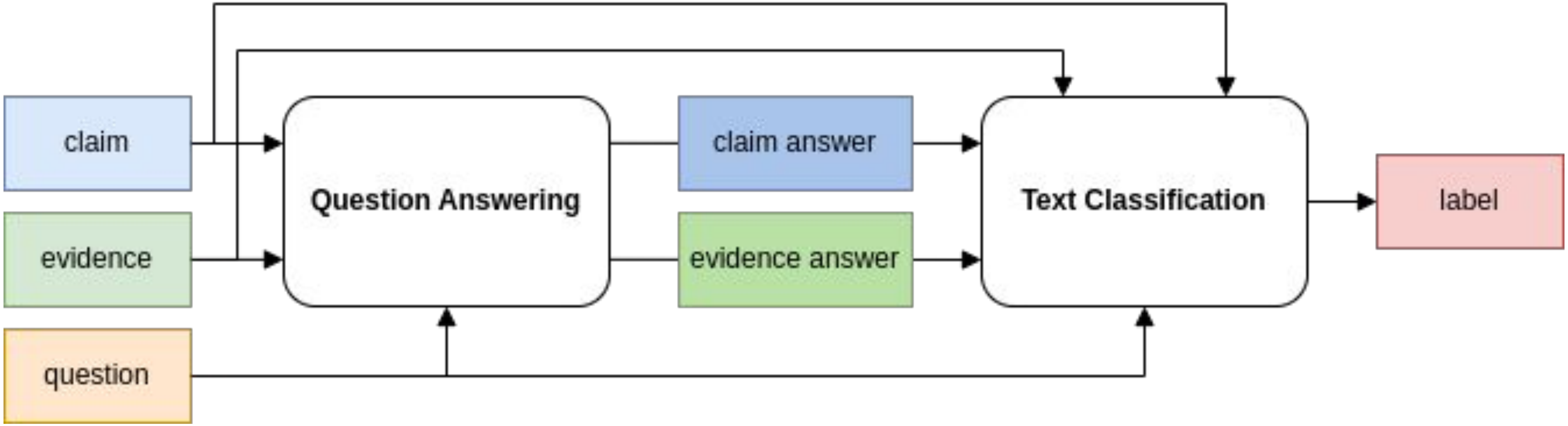
Our Solution: Pre-CoFactv3

ICL / Feature Extraction / Fine Tuning / Ensemble Learning

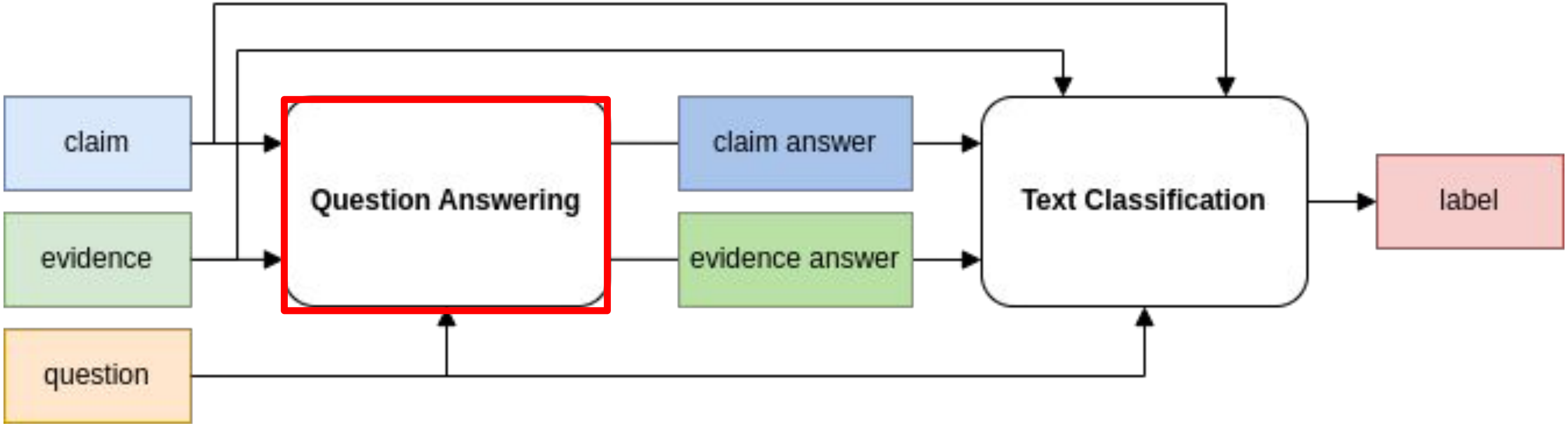
	Support	Neutral	Refute	Total
In-Context Learning Baseline	0.7500	0.2857	0.3333	0.4300
Human Baseline	0.5500	0.6000	0.1333	0.4400
Pre-CoFactv3	**0.8000**	**0.9133**	**0.8800**	**0.8644**

Method

Pre-CoFactv3 Overview



Question Answering



Fine-tuning Large Language Models (LLMs)

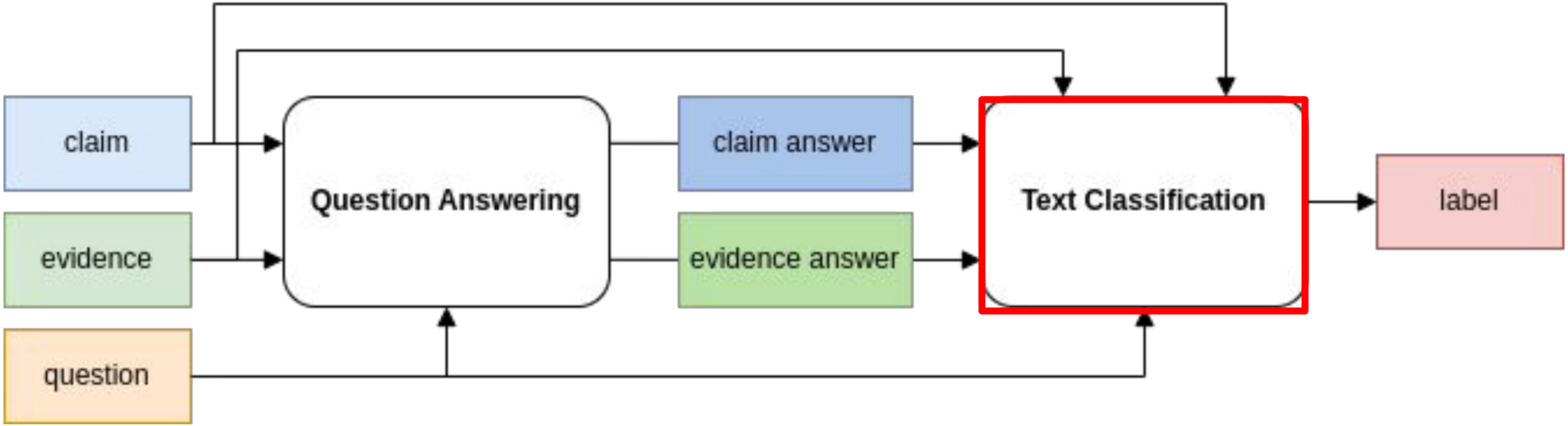
Claim Answer　　　　Claim　Question

$$index\ of\ \boxed{CA_{ij}}\ in\ C_i = LLM(\boxed{C_i}, \boxed{Q_{ij}})$$

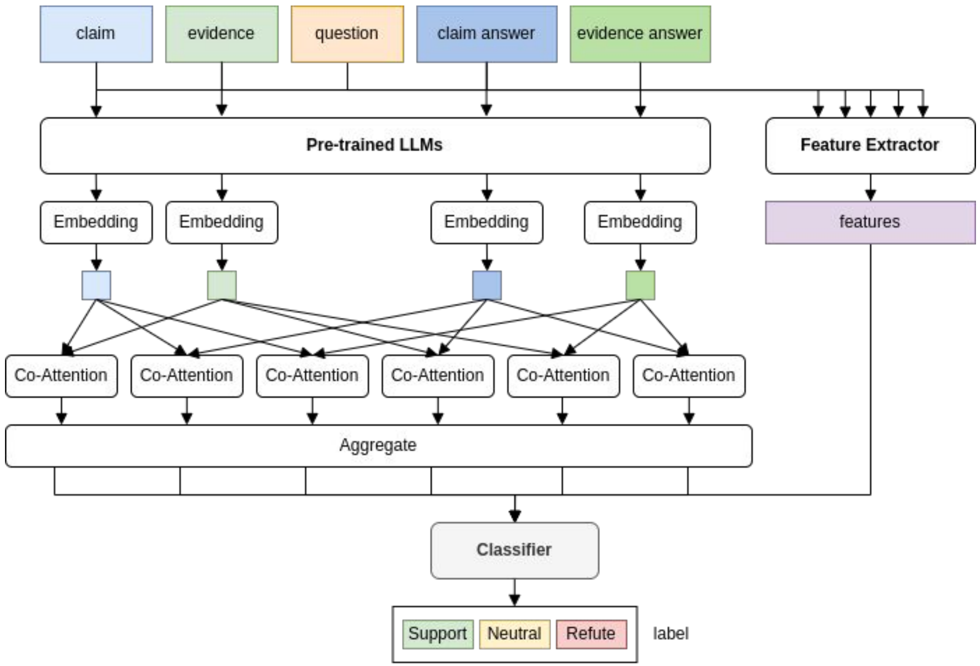
$$index\ of\ \boxed{EA_{ij}}\ in\ E_i = LLM(\boxed{E_i}, \boxed{Q_{ij}})$$

Evidence Answer　　　　Evidence

Text Classification

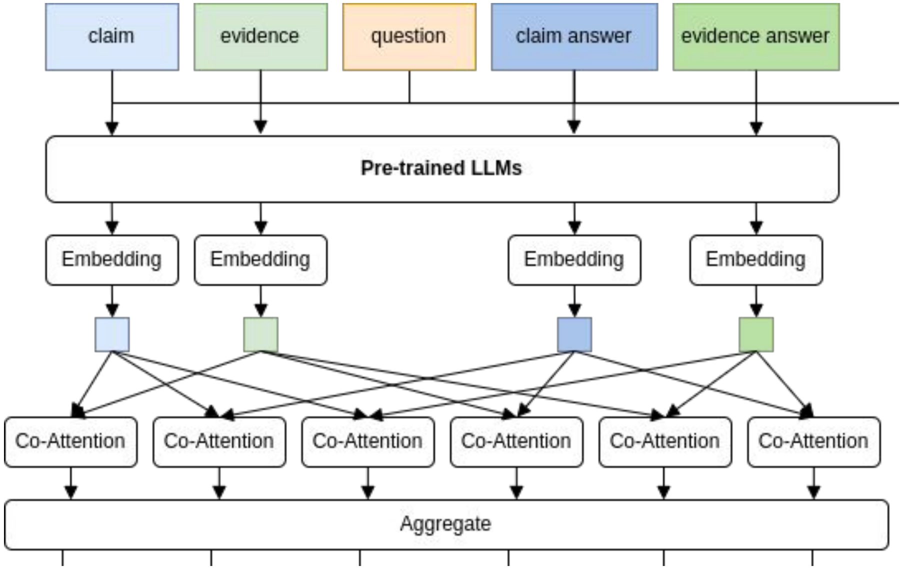


FakeNet

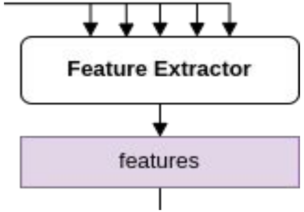

claim
evidence
question
claim answer
evidence answer
Pre-trained LLMs
Feature Extractor
Embedding
Embedding
Embedding
Embedding
features
Co-Attention
Co-Attention
Co-Attention
Co-Attention
Co-Attention
Co-Attention
Aggregate
Classifier
Support
Neutral
Refute
label

Pre-trained LLMs

1. Embedding by Pre-trained LLMs
2. Six Co-attentions
3. Mean Aggregation

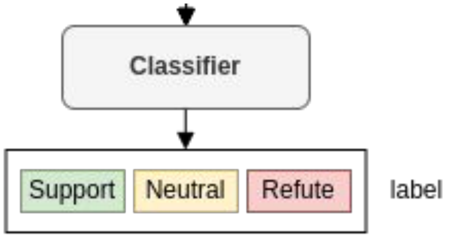


Feature Extractor



Type	Features		
Common Features in NLP	Character count	Word count	Count of capital characters
	Count of capital words	Count of punctuation	Count of words in quotes
	Sentence count	Count of unique words	Count of hashtags
	Count of mentions	Count of stopwords	
Similarity between Text Pair	SimCse	MPNet	The Fuzz
	TF-IDF	Rouge	

Classifier



Embeddings from the Pre-trained LLMs

Embeddings from the Feature Extractor

$$\hat{y}_i = softmax((\sigma((\boxed{E_{PLM}} + \boxed{E_{FE}})W^{Z1}))W^{Z2}),$$

Fine-tuning Large Language Models (LLMs)

$$\overset{\text{Claim}}{\boxed{C_i}} + \overset{\text{Evidence}}{\boxed{E_i}} + \overset{\text{Question}}{\boxed{Q_i}} + \overset{\text{Claim Answer}}{\boxed{CA_i}} + \overset{\text{Evidence Answer}}{\boxed{EA_i}}$$

$$I = \boxed{C_i} + \boxed{E_i} + \boxed{Q_i} + \boxed{CA_i} + \boxed{EA_i}$$

$$\hat{y}_i = LLM(I)$$

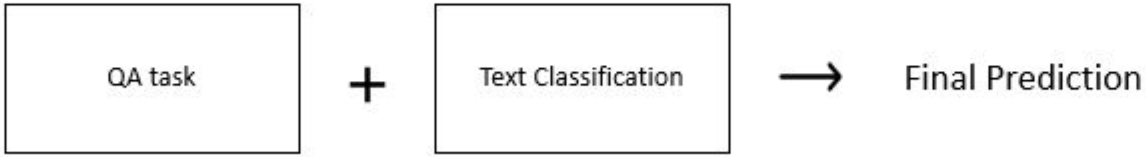
Ensemble

1. Weighted sum with labels
2. Power weighted sum with labels
3. Power weighted sum with two models
4. Power weighted sum with three models

Experiment

Experiment

- The official competition metric for Factify 3.0 involves 2 parts:



- A prediction is deemed correct if :
  - the BLEU score for QA task exceeds a predefined threshold
  - the predicted label of text classification is correct

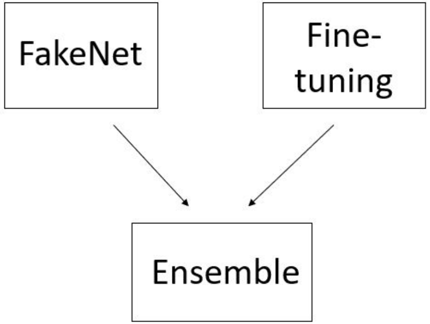
Question Answering

LLMs	Claim Answer (BLEU)	Evidence Answer (BLEU)	Average (BLEU)
1	0.3543	0.3006	0.3275
2	0.3586	0.3178	0.3382
3	0.5230	0.3361	0.4296
4	0.5248	**0.3963**	**0.4605**
5	**0.5323**	0.3518	0.4421
6	0.5268	0.3873	0.4571

Experiment result of Compare LLMs on the Question Answering task.

LLMS	Model	Fine-tuned dataset
1	Roberta-large	SQuAD 2.0
2	Deberta-large	SQuAD 2.0
3	Roberta-large	FACTIFY5WQA
4	Deberta-v3-large	FACTIFY5WQA
5	Roberta-large	SQuAD 2.0 + FACTIFY5WQA
6	Deberta-v3-large	SQuAD 2.0 + FACTIFY5WQA

Text Classification

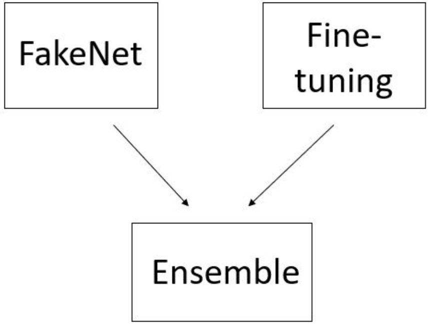


Method 1: FakeNet

Pre-trained LLMs	Epoch	Accuracy
bert-large-uncased [12]	20	0.7040
gpt2 [21]	30	0.6813
t5-large [22]	10	0.6991
microsoft/deberta-large [23]	20	0.7498
microsoft/deberta-xlarge [23]	20	0.7440
microsoft/deberta-v3-base [3]	15	0.7364
microsoft/deberta-v3-large [3]	**15**	**0.7542**

Experiment results of different Pre-trained LLMs in FakeNet.

Text Classification



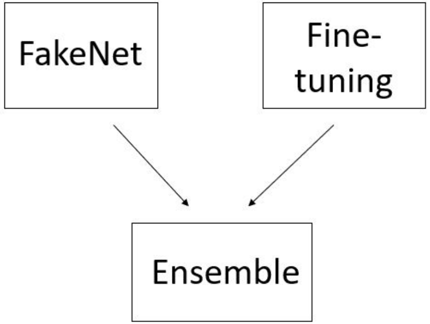
Method 2: Fine-tuning

Input	Claim Length	Evidence Length	Question Length	Evidence Answer Length	Claim Answer Length	Accuracy
text	100	1000	-	-	-	0.8044
text	400	4000	-	-	-	0.8396
text	800	8000	-	-	-	0.8462
text	1600	10000	-	-	-	**0.8502**
question + answer	-	-	50	50	100	0.6311
text + question + answer	100	1000	50	50	100	0.7849

The performance comparison between different alterations in the input and length.

Text Classification

Final Part: Ensemble



Ensemble Methods	Model 1	Model 2	Model 3	Accuracy
Weighted sum with labels	Fine-tuned LLM 1	Fine-tuned LLM 2	-	0.8564
Power weighted sum with labels	Fine-tuned LLM 1	Fine-tuned LLM 2	-	0.8587
Power weighted sum with two models	Fine-tuned LLM 1	Fine-tuned LLM 2	-	0.8609
Power weighted sum with three models	Fine-tuned LLM 1	Fine-tuned LLM 2	FakeNet	**0.8644**

The experimental results for the four ensemble methods

Result

Training result:

	Support	Neutral	Refute	Total
In-Context Learning Baseline	0.7500	0.2857	0.3333	0.4300
Human Baseline	0.5500	0.6000	0.1333	0.4400
Pre-CoFactv3	**0.8000**	**0.9133**	**0.8800**	**0.8644**

Accuracy of Human Baseline, In-Context Learning Baseline, and Pre-CoFactv3 with different labels.

Result

Testing result:

Submission	Question Answering	Text Classification	Testing Accuracy
1	Fine-tuned LLM	FakeNet	0.6880
2	Fine-tuned LLM	Fine-tuned LLM	**0.6956**
3	Fine-tuned LLM	Ensemble	0.6080

The testing accuracy of three submissions.

Team Name	Testing Accuracy (%)
Baseline	0.3422 (0%)
Jiankang Han	0.4547 (33%)
SRL_Fact_QA	0.4551 (33%)
Trifecta (Ours)	**0.6956 (103%)**

The leaderboard for the Factify 3.0 Workshop.

Conclusion

Conclusion

- Our team achieve first place in this workshop
- In the text classification part, we ensmeble two methods: FakeNet & Fine-tuning
- After conducting extensive ablation studies, we identified the optimal combination of methods and fine-tuning techniques

Thanks for listening!

Paper

Chiang Shang-Hsuan (andy10801@gmail.com)

Ming-Chih Lo (max230620089@gmail.com)

Lin-Wei Chao (williamchao.ii12@nycu.edu.tw)

GitHub

Appendix

Limitation

- The quality of the dataset may highly affect our model's performance
- The model is subject to certain contraints and may not accurately apply to real-world data
- Previously collected evidence may not confirm the news as it happens

Human & In-Context Learning Baseline

- In-Context Learning: dataset of 100 randomly selected from FACTIFY5WQA
- Human: dataset of 20 randomly selected from in-context learning dataset

	Support	Neutral	Refute	Total
In-Context Learning Baseline	0.7500	0.2857	0.3333	0.4300
Human Baseline	0.5500	0.6000	0.1333	0.4400
Pre-CoFactv3	**0.8000**	**0.9133**	**0.8800**	**0.8644**