Improving the fact-checking performance of language models by relying on their entailment ability

Anonymous ACL submission

Abstract

Automated fact-checking is a crucial task in this digital age. To verify a claim, current approaches majorly follow one of two strategies i.e. (i) relying on embedded knowledge of language models, and (ii) fine-tuning them with evidence pieces. While the former can make systems to hallucinate, the later have not been very successful till date. The primary reason behind this is that fact verification is a complex process. Language models have to parse through multiple pieces of evidence before making a prediction. Further, the evidence pieces often contradict each other. This makes the reasoning process even more complex. We proposed a simple yet effective approach where we relied on entailment and the generative ability of language models to produce "supporting" and "refuting" justifications (for the truthfulness of a claim). We trained language models based on these justifications and achieved superior results. Apart from that, we did a systematic comparison of different prompting and fine-tuning strategies, as it is currently lacking in the literature. Some of our observations are: (i) training language models with raw evidence sentences registered an improvement up to **8.20%** in macro-F1, over the best performing baseline for the RAW-FC dataset, (ii) similarly, training language models with prompted claim-evidence understanding (**TBE-2**) registered an improvement (with a margin up to **16.39%**) over the baselines for the same dataset, (iii) training language models with entailed justifications (**TBE-3**) outperformed the baselines by a huge margin (up to **28.57%** and **44.26%** for LIAR-RAW and RAW-FC, respectively). We have shared our code repository to reproduce the results.

1 Introduction:

The spread of misinformation on the internet has grown to be a pressing social issue. Its consequences have manifested across social, political and commercial domains (Mozur, 2018; Fisher et al., 2016; Allcott and Gentzkow, 2017; Burki, 2019; Aghababaeian et al., 2020). To counter the spread, institutional interventions doing manual fact-checking have gained momentum. For example, the International Fact-Checking Network (IFCN), started in 2015, works with over 170 fact-checking organisations and websites worldwide (as of July 2024 [1]). However, manually verifying facts and detecting misinformation is a slow and costly process. Human experts struggle to keep up with the pace of rapid spread. To overcome this, the research community have been trying to automate the process. Their interest can be gauged by the fact that more than twelve hundred research articles and more than fifty datasets were published on this topic (Alnabhan and Branco, 2024; Guo et al., 2022). A detailed list of such related works is reported in the Appendix A.

Doing automated fact-checking at scale is still an open challenge. Most of the proposed approaches relied on language models (Shu et al., 2022; Yang et al., 2022a; Yue et al., 2023; Choi and Ferrara, 2024a). Researchers have either (i) relied on their embedded knowledge or (ii) used evidence collected from various sources to predict the veracity. For example, Pan et al. (2021); Zeng and Gao (2023) reported the performance of *zero-shot* or *few-shot* prompting for this task. Dhuliawala et al. (2024) showcased that integrating reasoning steps through *chain-of-thought* prompting reduced model hallucinations. Others relied on external tools such as web search (Galitsky, 2025; Zhang and Gao, 2023) to retrieve the evidence and do fact verification. However, their accuracy is far behind to deploy them for practical purposes. The primary reason behind this is that fact verification is difficult because language models need to look at multiple pieces of information to decide what is true. These

[1] https://www.poynter.org/ifcn/

pieces often disagree with each other, which makes the reasoning process even harder. To counter this, Yue et al. (2024) segregated the retrieved evidence into 'supporting' and 'opposing' arguments, allowing the model to consider each perspective independently, and then they verified each claim using few-shot prompting. On a philosophically similar line, Wang et al. (2024a) introduced a module to extract top-k evidence sets stating the claim to be 'true' and 'false' based on attention scores. They produced 'true' and 'false' justifications by promoting language models based on evidence sentences, and further trained the language models to produce veracity by training LLMs. We, on the other hand, proposed a simple yet effective approach, where we relied on (i) the entailment ability of language models to classify if evidence sentences are supporting or refuting a given claim and (ii) the generative ability of language models to produce supporting and refuting justifications. A schematic diagram of our proposed approach is presented in Figure 17 (in the Appendix). We believe the simplicity of our approach will allow us to easily deploy our systems in the real-world. Apart from that, we also found that a systematic comparison of different prompting and fine-tuning strategies is lacking in the literature. To fill this research gap, we designed different experiments based on three philosophical questions. They are,

- ★ **R1:** *"How well do the language models perform when only raw evidence sentences along with the claim are available during training and inferencing?"*
- ★ **R2:** *"Does training and inferencing with prompted claim-evidence understanding improve the performance of the language models?"*
- ★ **R3:** *"Can training and inferencing with prompted entailed justifications improve the claim veracity prediction?"*

Our key contributions in this work are,

- We conducted three training-based (**TBE**) and four inference-based (**IBEs**) experiments along the line of the research questions. We gave three types of inputs, i.e, along with the claim (i) raw evidence sentences, (ii) overall claim-evidence understanding generated by language models, and (iii) entailed justifications generated by the language models.
- We conducted a detailed evaluation of model explanations. We considered the entailed justifications generated by language models as model explanations, and evaluated them using two strategies: (i) checking lexical overlap and semantic matching, and (ii) doing subjective evaluation by language models. While in the former, we used ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BERT-scores (Zhang et al., 2019), in the latter, we prompted VLLMs to check how informative, accurate, readable, objective, and logical the explanations are.
- We conducted an ablation study and a thorough error analysis of our models. In the ablation study, we removed individual supporting and refuting entailed justifications from input samples and trained the veracity prediction models again. This exercise illustrated the importance of individual entailed justifications. In linguistic analysis, on the other hand, we identified (i) the cases where our model succeeds and fails, and (ii) the possible reasons behind it.

Some of the interesting observations we got are,

- While prompting language models (**IBEs**) with raw evidence sentences couldn't outperform the baselines, training some language models with the same (**TBE-1**) registered an improvement up to **8.20%** in macro-F1 for the RAW-FC dataset. Training Language models with prompted claim-evidence understanding (**TBE-2**) registered an improvement (with a margin up to **16.39%**) over baselines for the RAW-FC dataset. Training language models with entailed justifications (**TBE-3**) outperformed the baselines by a large margin (up to **28.57%** and **44.26%** for LIAR-RAW and RAW-FC, respectively).
- Subjective evaluation by language models validated the correlations between veracity prediction and explanation quality in the best-performing models. For instance, Llama-generated explanations received highest informativeness, readability, objectivity, and logicality ratings, which correlates with the highest macro-f1 it got in veracity prediction.
- In the ablation study, we found that the macro-F1 score dropped when the best perform-

ing models were trained without supporting (**31.48**% ↓ in LIAR-RAW and **12.50**% ↓ in RAW-FC) and refuting justifications (**9.26**% ↓ in LIAR-RAW and **9.09**% ↓ in RAW-FC). Further, discarding both had the maximum adverse effect (**55.56**% ↓ in LIAR-RAW and **47.73**% ↓ in RAW-FC).

2 Dataset details:

In this section, we have reported the details of the datasets considered in our study. We have used the **LAIR-RAW** and **RAW-FC** datasets provided by Yang et al. (2022b) in our experiments. While each sample of LAIR-RAW is tagged with one of six labels, samples of RAW-FC is tagged with one of three labels. The list of labels and their distribution in respective datasets are reported in Table 1. The datasets are open-source, i.e., they are Apache 2.0 licensed. A detailed description of the individual datasets along with representative sample is reported in the Appendix B.

Dataset	Classes	Count
LAIR-RAW (Yang et al., 2022b)	True (T)	2,021
	Mostly-true(MT)	2,439
	Half-true (HT)	2,594
	Barely-true (BT)	2,057
	False (F)	2,466
	Pants-fire (PF)	1,013
	Total	12,590
RAW-FC (Yang et al., 2022b)	True (T)	695
	Half-true (HT)	671
	False (F)	646
	Total	2,012

Table 1: Datasets Statistics

3 Experiments:

In this section, we have reported the details of the experiments we conducted as part of this study. We conducted two types of experiments- (i) training-based (**TBE**), and (ii) inference-based (**IBE**). As their name suggests, in **TBE** approaches, we fine-tuned various language models to predict the veracity labels based on customised inputs. Particularly, (i) we fine-tuned large language models (LLMs) like RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019), and (ii) we fine-tuned very large language models (VLLMs) like Mistral (Jiang et al., 2023), Llama (AI@Meta, 2024), Gemma (Team et al., 2024), Qwen (Yang et al., 2024) and Falcon (Almazrouei et al., 2023) models using LoRA (Hu et al., 2022) and LoRA+ (Hayou et al., 2024) adapters. Note that, unlike LLMs, we can not directly fine-tune VLLMs due to their large parameter size and computational constraints. Similarly, in **IBE** approaches, we prompted the considered VLLMs to predict veracity labels without explicitly training them for it. We considered the same set of LLMs and VLLMs in all of our experiments reported in this study. The details of individual experiments in each category are reported in the subsequent subsections. We have also reported the current state-of-the-art models as the baselines.

3.1 Training Based Experiments (TBE):

In this section, we have reported the details of three training-based experiments we have considered in our study. Each of them is based on a unique research philosophy. The details of individual approaches are reported in the subsequent subsections.

3.1.1 TBE-1: Training based on raw-evidences:

In the first experiment, we tried to answer *"How well do the language models perform when raw evidence sentences are available during training?"*. To answer this, we fine-tuned the language models by giving claims and raw evidence sentences as input. For the samples which don't have associated raw-evidence sentences, we gave claims as the only input. Here, we restricted ourselves from fine-tuning LLMs like RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019), as the input length of many samples exceeded the maximum supported input size of these language models. Past work (Cheung and Lam, 2023) demonstrated the effectiveness of VLLM finetuning using evidence pieces from web search, whereas in our case, we used the gold evidence sentences given in the dataset. The emergence of adapter-based training for VLLMs allowed us to do this, as they can be trained with large input texts. To the best of our knowledge, we believe we are the first to fine-tune the VLLMs using raw evidence sentences. The approach we followed in **TBE-1** is illustrated in sub-figure (a) of Figure 1.

3.1.2 TBE-2: Training based on overall understanding:

In the second experiment, we tried to answer *"Does training with VLLM-generated claim-evidence understanding improve the performance of the language models?*. To conduct this experiment, we

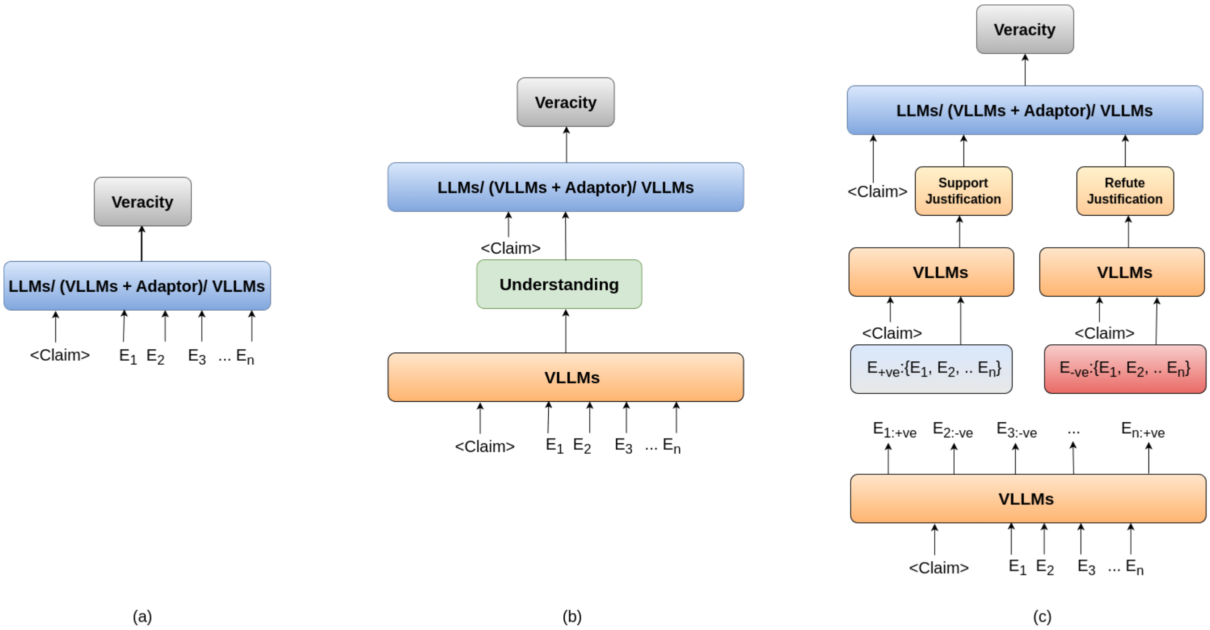


Figure 1: Illustration of steps we followed in different experiments. Sub-figure (a) presents the case where only raw evidence sentences and claims are given as input. This approach is used in **TBE-1** and **IBE-1**. In **TBE-1**, we trained the adapters (with VLLMs) using these inputs, whereas in **IBE-1**, we prompted VLLMs using zero-shot promptings. Sub-figure (b) shows the overall process of **TBE-2**, **IBE-2** and **IBE-3**. In **TBE-2**, the training of LLMs and VLLMs with adapters were done with the help of claim-evidence understandings generated by VLLMs. However, in **IBE-2** and **IBE-3**, we used zero-shot and CoT based prompting for final veracity prediction. Sub-figure (c) illustrates the overall experimental process of **TBE-3** and **IBE-4**. Here, first, we generate the entailment labels ("supporting" or "refuting") for individual evidence sentences for a given claim. We prompted the considered VLLMs to do the same. Subsequently, we clubbed the supporting and refuting evidence sentences and prompted VLLMs to generate justifications supporting and refuting the given claim. Lastly, based on the generated justifications, we generated the claim veracity by (i) training the LLMs and VLLMs with adapters as part of **TBE-3**, and (ii) prompting the VLLMs as a part of **IBE-3**.

first prompted the five considered VLLMs to generate their understanding of a given claim and its evidence sentence set. For the samples which don't have an associated evidence sentence, VLLMs generated their understandings based on the embedded knowledge. Based on the understanding, we fine-tuned the LLMs (such as RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019)) and trained the adapters (LoRA (Hu et al., 2022) and LoRA+ (Hayou et al., 2024)) with the considered VLLMs to produce the claim veracity. The detailed experimental process is illustrated in sub-figure (b) of Figure 1. Some of the prompt samples are presented in the appendix (Figure 18).

3.1.3 TBE-3: Training based on entailment understanding:

In the third experiment, we tried to answer *"Can training with VLLM-generated entailing justifications improve the claim veracity prediction?"*. To conduct this experiment, we followed a three-step approach. In the first step, we prompted the considered five VLLMs to classify if the evidence sentences are "supporting" or "refuting" a given claim. Our approach is inspired by the L-Defense model proposed by Wang et al. (2024a), where the authors used an attention mechanism to distinguish supporting and refuting evidence sentences. We, however, used the VLLMs for the same. In the second step, we prompted the language models to generate supporting and refuting *justifications* based on the classified evidence sentences. For the cases where claims don't have any supporting or refuting evidence sentences, VLLMs generated justifications based on their embedded knowledge. We separately passed the supporting and refuting evidence sentences associated with a claim in the prompts to generate them. Finally, based on the claims and the generated justifications, we (i) fine-tuned the LLMs (such as RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019)), and (ii) fine-tuned the adapters (LoRA (Hu et al., 2022) and LoRA+ (Hayou et al., 2024)) with considered VLLMs to generate the veracity labels. The detailed approach is illustrated in sub-figure (c) of Figure 1. Some prompt samples we used at each step are illustrated in Figure 19, Figure 20, and Figure 21 in the appendix.

3.2 Inference Based Experiments (IBE):

We have experimented with four types of inference-based approaches, each based on a unique prompting philosophy. They are (i) *zero-shot prompting* (**IBE-1**), where we have prompted five considered VLLMs to predict the veracity label given a claim and its associated evidence sentences, (ii) *zero-shot prompting with overall understanding* (**IBE-2**), where, first, we prompted the five considered VLLMs to generate the claim-evidence understanding and then, we prompted them again to with the understanding to predict the veracity labels, (iii) *CoT prompting with overall understanding* (**IBE-3**), where we followed similar steps as mentioned in **IBE-2**, except, additionally, we asked the VLLMs to generate step-by-step reasoning behind their predictions, and (iv) *prompting based on entailment* (**IBE-4**), where we prompted the VLLMs to predict veracity labels based on entailed justifications. Due to space constraints, we have reported the details of individual approaches in the appendix. For the cases where claims don't have associated evidence sentences, we gave the claim as only input (in **TBE-1**), generated understandings and justifications based on the embedded knowledge of VLLMs.

3.3 Baselines:

In this section, we have reported the previously proposed best-performing models as the baselines. Particularly, we compared our models with the performances of HiSS (Zhang and Gao, 2023), FactLLaMa (Cheung and Lam, 2023), RAFTS (Yue et al., 2024), and L-Defence (Wang et al., 2024a) models. Out of them, HiSS (Zhang and Gao, 2023) and FactLLaMa (Cheung and Lam, 2023) retrieve evidence from external sources, while RAFTS (Yue et al., 2024) employs a coarse-to-fine retrieval technique to extract evidence directly from the dataset. In contrast, L-Defense (Wang et al., 2024a) used relevant evidence without additional retrieval. The previously reported performance of these models on LIAR-RAW and RAW-FC datasets are presented in Table 2. A detailed description of the individual models is presented in the Appendix C.5.

Method	**LIAR-RAW**			**RAWFC**		
	MP	**MR**	**MF1**	**MP**	**MR**	**MF1**
HiSS	0.46	0.31	0.37	0.53	0.54	0.53
FactLLaMA	0.32	0.32	0.30	0.56	0.55	0.55
RAFTS	**0.47**	**0.37**	**0.42**	**0.62**	0.52	0.57
L-Defense						
-ChatGPT	0.30	0.32	0.30	0.61	**0.61**	**0.61**
-Llama2	0.31 (0.29†)	0.31 (0.29†)	0.31 (0.29†)	0.61 (0.56†)	0.60 (0.56†)	0.60 (0.56†)

Table 2: Performance of the considered baseline methods on the LIAR-RAW and RAWFC datasets. Notation: our reproduced results are marked as '†'.

Dataset (→)	**LIAR-RAW**			**RAW-FC**		
	TBE-1			TBE-1		
Method (↓)	MP	MR	MF1	MP	MR	MF1
LoRA						
-Mistral	**0.44** (±0.02)	0.29 (±0.01)	0.27 (±0.01)	0.69 (±0.01)	0.65 (±0.00)	0.65 (±0.01)
-Llama	0.34 (±0.01)	**0.30** (±0.02)	**0.30** (±0.00)	0.68 (±0.01)	0.64 (±0.02)	0.65 (±0.02)
-Gemma	0.27 (±0.02)	0.25 (±0.03)	0.23 (±0.04)	0.60 (±0.05)	0.58 (±0.04)	0.57 (±0.06)
-Qwen	0.37 (±0.02)	0.29 (±0.02)	0.29 (±0.04)	0.67 (±0.03)	**0.66** (±0.03)	**0.66** (±0.03)
-Falcon	0.39 (±0.01)	0.28 (±0.01)	0.26 (±0.01)	0.59 (±0.03)	0.58 (±0.05)	0.54 (±0.06)
LoRA+						
-Mistral	0.40 (±0.03)	0.27 (±0.02)	0.25 (±0.01)	0.53 (±0.05)	0.56 (±0.03)	0.55 (±0.02)
-Llama	0.34 (±0.01)	0.29 (±0.01)	0.29 (±0.01)	0.67 (±0.01)	0.64 (±0.01)	0.65 (±0.01)
-Gemma	0.27 (±0.02)	0.23 (±0.02)	0.22 (±0.03)	0.61 (±0.03)	0.57 (±0.02)	0.57 (±0.03)
-Qwen	0.36 (±0.01)	0.29 (±0.01)	0.29 (±0.02)	**0.70** (±0.02)	0.65 (±0.02)	0.65 (±0.02)
-Falcon	0.37 (±0.02)	**0.30** (±0.01)	0.29 (±0.02)	0.64 (±0.03)	0.62 (±0.03)	0.63 (±0.03)

Table 3: Performance of LoRA and LoRA+ models on TBE-1 (LIAR-RAW and RAW-FC datasets).

4 Results and Discussion:

In this section, we have reported the results from all of our experiments. While we used macro-precision, macro-recall and macro-F1 scores to evaluate the veracity, ROUGE, BLEU and BERT-score, along with subjective evaluation by VLLMs, were made to evaluate the explanations. A detailed description of evaluation metrics is reported in sub-section C.6.4 in the appendix.

4.1 Observations from TBE-1:

We reported the performance of various models for **TBE-1**, i.e. training with raw evidence sentences as input in Table 3. We found that LoRA adapter trained with Llama VLLM resulted in the highest F1 score of **0.30** for the LIAR-RAW dataset. While this performance is comparable to the performance of baseline models L-Defense ($F1$: **0.31**) and

FactLLaMA ($F1$: **0.30**), it does not surpass the performance of other baselines, i.e. HiSS ($F1$: **0.37**) and RAFTS ($F1$: **0.42**). However, for the RAWFC dataset, we found that many models, such as Mistral with LoRA ($F1$: **0.65**, $\sim$ **6.56**% $\uparrow$), Llama with LoRA ($F1$: **0.65**, $\sim$ **6.56**% $\uparrow$), Qwen with LoRA ($F1$: **0.66**, $\sim$ **8.20**% $\uparrow$), Llama with LoRA+ ($F1$: **0.65**, $\sim$ **6.56**% $\uparrow$), Qwen with LoRA+ ($F1$: **0.65**, $\sim$ **6.56**% $\uparrow$) and Falcon with LoRA+ ($F1$: **0.63**, $\sim$ **3.28**% $\uparrow$), outperform the best baseline performance ($F1$: **0.61**). Here, $\sim x\% \uparrow$ denotes the relative percentage improvement compared to the best baseline score. Qwen trained with LoRA resulted in the overall highest F1-score (**0.66**, $\sim$ **8.20**% $\uparrow$). We have reported some additional observations in the appendix section D.1.

4.2 Observations from TBE-2:

We reported the performance of various models for **TBE-2**, i.e. training with claim-evidence understanding as input in table 4. We found that Mistral trained with LoRA resulted in the highest macro-F1 (**0.32**) for the LIAR-RAW dataset. While its performance surpasses the performance of baselines FactLLaMA ($MF1$: **0.30**) and L-Defense ($MF1$: **0.31**), it is still behind HiSS ($MF1$: **0.37**) and RAFTS ($MF1$: **0.42**). However, for the RAW-FC dataset, we observed that many models, such as XLNet fine-tuned on Llama based understandings ($MF1$: **0.62**, $\sim$ **1.64**% $\uparrow$), and Llama trained (with Llama understandings) with LoRA+ ($MF1$: **0.71**, $\sim$ **16.39**% $\uparrow$) outperformed the best reported macro-F1 score by the baselines ($MF1$: **0.61**). Llama trained (with Llama understandings) with LoRA+ adapter resulted in the highest overall performance ($MF1$: **0.71**, $\sim$ **16.39**% $\uparrow$). We have reported some additional observations in the section D.2 of the appendix due to space constraints.

4.3 Observations from TBE-3:

We reported the performance of various models for **TBE-3**, i.e. training with entailed justifications generated by VLLMs as input in table 4. We found that XLNet fine-tuned on Llama based entailed justification resulted in the highest macro-F1 (**0.54**, $\sim$ **28.57**% $\uparrow$) for the LIAR-RAW dataset. Many models, such as RoBERTa fine-tuned with Mistral ($MF1$: **0.47**, $\sim$ **16.39**% $\uparrow$), Llama ($MF1$: **0.52**, $\sim$ **23.81**% $\uparrow$), Gemma ($MF1$: **0.48**, $\sim$ **14.29**% $\uparrow$), Qwen ($MF1$: **0.46**, $\sim$ **9.52**% $\uparrow$), and Falcon ($MF1$: **0.44**, $\sim$ **4.76**% $\uparrow$) based entailed justifications, XLNet fine-tuned with Mistral ($MF1$: **0.47**, $\sim$ **16.39**% $\uparrow$), Llama ($MF1$: **0.54**, $\sim$ **28.57**% $\uparrow$), Qwen ($MF1$: **0.48**, $\sim$ **14.29**% $\uparrow$), and Falcon ($MF1$: **0.44**, $\sim$ **4.76**% $\uparrow$) based entailed justifications, and Llama trained (with Llama justifications) with LoRA+ adapter ($MF1$: **0.49**, $\sim$ **16.67**% $\uparrow$) surpassed the best reported macro-F1 score of baselines ($MF1$: **0.42**). In a similar manner, for the RAW-FC dataset, we observed that many models, such as such as RoBERTa fine-tuned with Mistral ($MF1$: **0.83**, $\sim$ **36.07**% $\uparrow$), Llama ($MF1$: **0.88**, $\sim$ **44.26**% $\uparrow$), Qwen ($MF1$: **0.71**, $\sim$ **16.39**% $\uparrow$), and Falcon ($MF1$: **0.64**, $\sim$ **4.91**% $\uparrow$) based entailed justifications, XLNet fine-tuned with Mistral ($MF1$: **0.82**, $\sim$ **34.42**% $\uparrow$), Llama ($MF1$: **0.87**, $\sim$ **42.62**% $\uparrow$), Qwen ($MF1$: **0.70**, $\sim$ **14.75**% $\uparrow$), and Falcon ($MF1$: **0.74**, $\sim$ **21.31**% $\uparrow$) based entailed justifications, and Llama trained (with Llama justifications) with LoRA+ adapter ($MF1$: **0.83**, $\sim$ **36.07**% $\uparrow$) outperformed the best reported macro-F1 score by the baselines ($MF1$: **0.61**). RoBERTa fine-tuned with Mistral based entailed justification achieved the highest overall performance ($MF1$: **0.88**, $\sim$ **44.26**% $\uparrow$). We have reported some additional observations in the Appendix section D.3.

4.4 Observations from IBEs':

We reported the performance of various models in **IBEs** in Table 11. For LIAR-RAW dataset, Mistral achieved the highest macro-F1 (**0.22**) in IBE-1. While in IBE-2, Llama achieved the highest performance ($MF1$: **0.22**), Mistral, Llama, Qwen and Falcon ($MF1$: **0.21**) gave the highest performance in IBE-3. In IBE-4, Mistral ($MF1$: **0.14**) and Falcon ($MF1$: **0.14**) got the highest performance. None of the macro-F1 could surpass the baseline performances. In contrast, for RAW-FC dataset, Llama ($MF1$: **0.62**) achieved the highest performace in IBE-2, surpassing the highest baseline performance ($MF1$: **0.61**). While Qwen attained the highest performance in IBE-1 ($MF1$: **0.59**), surpassing the baselines HiSS ($MF1$: **0.53**), FactL-LaMa ($MF1$: **0.55**) and RAFTS ($MF1$: **0.57**), it is still behind the performance of L-Defense ($MF1$: **0.61**). Similarly, Qwen got the highest performance in IBE-3 ($MF1$: **0.52**). In IBE-4, Mistral and Qwen achieved the highest macro-F1 of **0.43**, which is far behind all baselines. We have reported some additional observations in the Appendix (section D.4)

Dataset (→)	LIAR-RAW						RAW-FC					
Method (↓)	TBE-2			TBE-3			TBE-2			TBE-3		
	MP	MR	MF1	MP	MR	MF1	MP	MR	MF1	MP	MR	MF1
FINE-TUNING												
-RoBERTa-L$_{Mistral}$	0.28 (±0.01)	0.26 (±0.01)	0.26 (±0.01)	0.48 (±0.01)	0.47 (±0.00)	0.47 (±0.01)	0.51 (±0.01)	0.50 (±0.01)	0.50 (±0.01)	0.83 (±0.00)	0.82 (±0.01)	0.83 (±0.01)
-RoBERTa-L$_{Llama}$	0.27 (±0.02)	0.28 (±0.01)	0.25 (±0.02)	0.53 (±0.01)	0.53 (±0.01)	0.52 (±0.01)	0.50 (±0.01)	0.50 (±0.01)	0.49 (±0.01)	**0.88** (±0.01)	**0.88** (±0.01)	**0.88** (±0.01)
-RoBERTa-L$_{Gemma}$	0.28 (±0.02)	0.27 (±0.01)	0.27 (±0.01)	0.49 (±0.01)	0.50 (±0.02)	0.48 (±0.02)	0.51 (±0.04)	0.51 (±0.03)	0.50 (±0.04)	0.50 (±0.02)	0.49 (±0.01)	0.49 (±0.02)
-RoBERTa-L$_{Qwen}$	0.30 (±0.01)	0.28 (±0.02)	0.28 (±0.01)	0.48 (±0.01)	0.47 (±0.02)	0.46 (±0.01)	0.51 (±0.01)	0.49 (±0.01)	0.48 (±0.02)	0.73 (±0.02)	0.71 (±0.02)	0.71 (±0.02)
-RoBERTa-L$_{Falcon}$	0.29 (±0.02)	0.28 (±0.02)	0.27 (±0.01)	0.49 (±0.01)	0.43 (±0.02)	0.44 (±0.01)	0.50 (±0.00)	0.48 (±0.01)	0.48 (±0.01)	0.65 (±0.02)	0.64 (±0.02)	0.64 (±0.02)
-XLNet-L$_{Mistral}$	0.29 (±0.02)	0.28 (±0.01)	0.28 (±0.01)	0.49 (±0.00)	0.47 (±0.01)	0.47 (±0.01)	0.62 (±0.02)	0.61 (±0.01)	0.61 (±0.02)	0.83 (±0.01)	0.82 (±0.01)	0.82 (±0.01)
-XLNet-L$_{Llama}$	0.31 (±0.02)	0.29 (±0.01)	0.29 (±0.01)	**0.55** (±0.01)	**0.54** (±0.02)	**0.54** (±0.01)	0.63 (±0.03)	0.63 (±0.03)	0.62 (±0.03)	**0.88** (±0.01)	**0.88** (±0.01)	0.87 (±0.01)
-XLNet-L$_{Gemma}$	0.26 (±0.02)	0.26 (±0.02)	0.25 (±0.02)	0.47 (±0.04)	0.43 (±0.05)	0.42 (±0.09)	0.51 (±0.01)	0.50 (±0.01)	0.50 (±0.01)	0.47 (±0.02)	0.47 (±0.02)	0.46 (±0.02)
-XLNet-L$_{Qwen}$	0.29 (±0.02)	0.29 (±0.02)	0.28 (±0.02)	0.50 (±0.01)	0.48 (±0.01)	0.48 (±0.01)	0.60 (±0.02)	0.58 (±0.01)	0.58 (±0.01)	0.70 (±0.03)	0.70 (±0.03)	0.70 (±0.04)
-XLNet-L$_{Falcon}$	0.26 (±0.03)	0.26 (±0.01)	0.24 (±0.03)	0.45 (±0.01)	0.44 (±0.02)	0.44 (±0.01)	0.61 (±0.02)	0.60 (±0.01)	0.60 (±0.01)	0.76 (±0.01)	0.75 (±0.01)	0.74 (±0.01)
LORA												
-Mistral	**0.36** (±0.01)	**0.32** (±0.01)	**0.32** (±0.01)	0.35 (±0.04)	0.34 (±0.03)	0.29 (±0.03)	0.61 (±0.04)	0.60 (±0.04)	0.60 (±0.04)	0.69 (±0.05)	0.59 (±0.05)	0.58 (±0.05)
-Llama	0.31 (±0.04)	0.27 (±0.03)	0.25 (±0.04)	0.36 (±0.03)	0.32 (±0.03)	0.30 (±0.04)	0.48 (±0.06)	0.48 (±0.03)	0.46 (±0.02)	0.63 (±0.03)	0.62 (±0.03)	0.61 (±0.03)
-Gemma	0.21 (±0.01)	0.22 (±0.01)	0.18 (±0.01)	0.26 (±0.03)	0.23 (±0.03)	0.20 (±0.02)	0.47 (±0.03)	0.45 (±0.03)	0.40 (±0.05)	0.36 (±0.02)	0.34 (±0.01)	0.30 (±0.02)
-Qwen	0.32 (±0.09)	0.27 (±0.02)	0.21 (±0.02)	0.40 (±0.07)	0.36 (±0.07)	0.32 (±0.05)	0.56 (±0.07)	0.54 (±0.0)	0.53 (±0.03)	0.48 (±0.03)	0.46 (±0.03)	0.46 (±0.03)
-Falcon	0.28 (±0.01)	0.26 (±0.00)	0.23 (±0.01)	0.27 (±0.03)	0.24 (±0.03)	0.21 (±0.02)	0.56 (±0.03)	0.57 (±0.01)	0.53 (±0.00)	0.43 (±0.01)	0.0.40 (±0.03)	0.41 (±0.03)
LORA+												
-Mistral	0.33 (±0.01)	0.30 (±0.01)	0.30 (±0.01)	0.30 (±0.07)	0.32 (±0.02)	0.27 (±0.03)	0.66 (±0.05)	0.60 (±0.05)	0.60 (±0.05)	0.52 (±0.06)	0.48 (±0.04)	0.46 (±0.04)
-Llama	0.27 (±0.01)	0.24 (±0.02)	0.23 (±0.02)	0.49 (±0.03)	0.50 (±0.02)	0.49 (±0.02)	**0.73** (±0.02)	**0.71** (±0.02)	**0.71** (±0.02)	0.84 (±0.01)	0.82 (±0.03)	0.82 (±0.03)
-Gemma	0.23 (±0.03)	0.21 (±0.02)	0.20 (±0.02)	0.34 (±0.04)	0.32 (±0.05)	029 (±0.06)	0.53 (±0.02)	0.51 (±0.02)	0.51 (±0.04)	0.52 (±0.03)	0.48 (±0.04)	0.46 (±0.01)
-Qwen	0.27 (±0.04)	0.27 (±0.03)	0.24 (±0.04)	0.33 (±0.03)	0.34 (±0.03)	0.29 (±0.03)	0.57 (±0.07)	0.48 (±0.03)	0.43 (±0.01)	0.52 (±0.06)	0.49 (±0.09)	0.47 (±0.09))
-Falcon	0.33 (±0.00)	0.28 (±0.01)	0.28 (±0.02)	0.39 (±0.02)	0.39 (±0.03)	0.36 (±0.04)	0.55 (±0.03)	0.56 (±0.02)	0.53 (±0.03)	0.58 (±0.01)	0.52 (±0.02)	0.50 (±0.03)

Table 4: Performance of claim veracity prediction using gold evidences. **Green** and Blue indicates best and second-best performance, respectively.

4.5 Comparing the performances across TBEs and IBEs:

In this section, we have reported our findings by comparing the best-performing models across all TBEs and IBEs. The comparison is also illustrated in Figure 22, Figure 23 and Figure 24 in the Appendix D.5. For the LIAR-RAW dataset, we found that none of the models from any IBEs, TBE-1 and TBE-2 ($MF1$: **0.14** - **0.32**) could surpass the highest F1 achieved in baselines ($MF1$: **0.42**). However, we observed that TBE-1 and TBE-2 ($MF1$: **0.30** - **0.32**, $\sim$ **45.45**% ↑) outperformed IBEs ($MF1$: **0.22**) in terms of the highest reported macro-F1. Further, we also found that many models from TBE-3 ($MF1$: **0.44** - **0.54**, $\sim$ **68.75**% ↑) outperform all of the models of IBEs ($MF1$: **0.14** - **0.22**), TBE-1 ($MF1$: **0.30**) and TBE-2 ($MF1$: **0.32**). For the RAW-FC dataset, we observed a similar comparative performance. However, here, the highest reported macro-F1 by IBE-2 (**0.62**) is comparable to the best-performing baseline ($MF1$: **0.61**). Similar to the LIAR-RAW dataset, here also we observed that (i) many TBE-1 and TBE-2 models ($MF1$: **0.62** - **0.71**, $\sim$ **14.52**% ↑) outperform IBEs (**0.62**) in terms of macro-F1, and (ii) many

Dataset (→)	LIAR-RAW												RAW-FC											
Method (↓)	IBE-1			IBE-2			IBE-3			IBE-4			IBE-1			IBE-2			IBE-3			IBE-4		
	MP	MR	MF1	MP	MR	MF1	MP	MR	MF1	MP	MR	MF1	MP	MR	MF1	MP	MR	MF1	MP	MR	MF1	MP	MR	MF1
PROMPTING																								
-Mistral	0.24	**0.25**	**0.22**	0.22	**0.23**	0.20	**0.40**	**0.23**	**0.21**	**0.30**	**0.17**	**0.14**	0.54	0.54	0.53	0.58	0.58	0.58	0.50	0.52	0.45	0.45	**0.46**	**0.43**
-Llama	0.24	0.23	0.20	**0.30**	**0.23**	**0.22**	0.26	0.22	**0.21**	0.22	0.13	0.13	0.56	0.56	0.54	**0.62**	**0.63**	**0.62**	0.45	0.47	0.49	0.42	0.39	0.35
-Gemma	0.24	0.21	0.19	0.16	0.16	0.13	0.27	0.19	0.16	0.09	0.16	0.11	0.40	0.40	0.40	0.41	0.38	0.38	0.45	0.41	0.40	0.27	0.31	0.24
-Qwen	**0.27**	0.23	0.20	0.25	**0.23**	0.20	0.24	0.22	**0.21**	0.13	0.16	0.13	**0.61**	**0.58**	**0.59**	0.58	0.57	0.57	0.57	**0.54**	**0.52**	**0.50**	**0.46**	**0.43**
-Falcon	0.24	0.23	0.20	0.22	0.22	0.20	0.24	**0.23**	**0.21**	0.16	**0.17**	**0.14**	0.56	0.57	0.54	0.60	0.59	0.57	**0.60**	0.52	0.48	0.40	0.38	0.37

Table 5: Performance of Prompting methods across IBE-1, IBE-2, and IBE-3 settings for LIAR-RAW and RAW-FC datasets.

TBE-3 models ($MF1$: **0.64** - **0.88**, $\sim$ **23.94**% ↑) outperform all of the models of IBEs ($MF1$: **0.43** - **0.62**), TBE-1 ($MF1$: **0.65** - **0.66**) and TBE-2 ($MF1$: **0.62** - **0.71**). The whole comparison points to the fact that training the LLMs with VLLM-generated entailed-justifications can improve the performance of veracity prediction. We have reported some additional observations in the section D.5 of the appendix due to space constraints. Further, we have also presented the confusion metrics of best-performing TBE and IBE models for LIAR-RAW and RAW-FC datasets in Figure 37 and 38 respectively.

4.6 Insights from evaluation of explanations:

In this section, we reported our findings from evaluating model explanations. We obtained them by concatenating the supporting and refuting justifications. The details of the metrics and evaluation scores are provided in Appendix C.6.4 and Appendix D.6, respectively. We observed that lexical overlap measures like ROUGE (R_1, R_2, R_L) provided mixed signals. For instance, while Falcon generated explanations showed higher unigram-based overlaps (R_1: **0.23** for LIAR-RAW and R_1: **0.40** for RAW-FC), Mistral generated explanations got the highest overlap in longest common subsequences (R_L: **0.14** for LIAR-RAW and R_L: **0.18** for RAW-FC). However, all models scored consistently low on the BLEU score and the BERT score. On the other hand, subjective evaluations indicated that Llama-generated explanations were rated highest by the majority of language models for dimensions like informativeness, readability, objectivity, and logicality for both datasets. These high subjective ratings seem to be correlated with better veracity prediction performance, as RoBERTa and XLNet performed well upon taking Llama explanations. A more detailed discussion on the observations of 'evaluation of explanations' is presented in Appendix D.6.

5 Conclusion:

- Training language models with VLLM entailed justifications surpassed the baseline macro-F1 scores substantially with an improvement of **28.57%** and **44.26%** for LIAR-RAW and RAW-FC, respectively. The approach of training with claim-evidence understanding (TBE-2) secured the second spot, with an increment of **16.39%** in the RAW-FC dataset compared to the best baseline macro-F1. In contrast, the inference-based methods (IBEs) were unable to understand the justifications generated from VLLM and performed consistently poorly.

- While lexical overlap and semantic matching methods showed no definite pattern, the subjective evaluation of model explanations by VLLMs found that Llama generated model explanations are more informative, readable, objective, and logical. It correlates with the superior performance reported by XLNet and RoBERTa (in **TBE-3**) for veracity prediction as they took Llama generated justifications as input.

- The role of VLLM entailed justification as a second step in TBE-3 is justified in the ablation study (see Appendix E) where we observed that removing supporting and refuting justification adversarially affected the scores.

- In the linguistic analysis of model explanations, we found that LLM could attend to supporting and refuting keywords and factual pieces of information for samples labelled with 'true' and 'false' veracity samples. However, for samples with other veracity labels LLM attention seem to be scattered. A detailed discussion is presented in Appendix F due to space constraints.

6 Limitations:

In this section, we reported the limitations of our work.

- We restricted our experiments to using only open-source language models, for reproducibility and resource constraints. However, expanding these experiments to commercial language models will further generalize the idea of claim-evidence entailment.
- Due to our limited linguistic expertise, we restricted our experiments to an English-language setup only. In the future, our hypothesis can also be tested in other languages.
- In our work, we assumed a closed-domain fact-checking setup for reproducibility. In future, one can consider open-domain fact-checking, i.e., retrieve evidence from an external source and test our hypothesis for generalisation.
- Due to space constraints, we restricted our experiments to the task of fact-checking and utilised two popular datasets. A wider testing of our hypothesis on various other datasets could help us know the applicability of our idea of utilising entailment.
- Due to resource constraints, we could not evaluate the model-generated explanations manually. One can extend the study by integrating human evaluation of the explanations.

References

Hamidreza Aghababaeian, Lara Hamdanieh, and Abbas Ostadtaghizadeh. 2020. Alcohol intake in an attempt to fight covid-19: A medical myth in iran. *Alcohol*, 88:29–32.

AI@Meta. 2024. Llama 3. `https://github.com/meta-llama/llama3`. Accessed: 2025-04-21.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023a. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023b. Multimodal automated fact-checking: A survey. *arXiv preprint arXiv:2305.13507*.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the first workshop on fact extraction and verification (FEVER)*, pages 85–90.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models.

Mohammad Q Alnabhan and Paula Branco. 2024. Fake news detection using deep learning: A systematic literature review. *IEEE Access*.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. 2019. Deepfake video detection through optical flow based cnn. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1205–1207.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Talha Burki. 2019. Vaccine misinformation and social media. *The Lancet Digital Health*, 1(6):e258–e259.

Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2025. Multimodal fact-checking with vision language models: A probing classifier based solution with embedding strategies. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4622–4633, Abu Dhabi, UAE. Association for Computational Linguistics.

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: detecting and preventing clickbaits in online news media. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '16, page 9–16. IEEE Press.

Tsun-Hin Cheung and Kin-Man Lam. 2023. Factllama: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 846–853. IEEE.

Eun Cheol Choi and Emilio Ferrara. 2024a. Fact-gpt: Fact-checking augmentation via claim matching with llms. In *Companion Proceedings of the ACM Web Conference 2024*, pages 883–886.

Eun Cheol Choi and Emilio Ferrara. 2024b. Fact-gpt: Fact-checking augmentation via claim matching with llms. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 883–886, New York, NY, USA. Association for Computing Machinery.

IDO DAGAN, BILL DOLAN, BERNARDO MAGNINI, and DAN ROTH. 2010. Recognizing textual entailment: Rational, evaluation and approaches – erratum. *Natural Language Engineering*, 16(1):105–105.

Shih-Chieh Dai, Yi-Li Hsu, Aiping Xiong, and Lun-Wei Ku. 2022. Ask to know more: Generating counterfactual explanations for fake claims. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2800–2810, New York, NY, USA. Association for Computing Machinery.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

John Dougrez-Lewis, Elena Kochkina, Miguel Arana-Catania, Maria Liakata, and Yulan He. 2022. PHEMEPlus: Enriching social media rumour verification with external evidence. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 49–58, Dublin, Ireland. Association for Computational Linguistics.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2024. Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6679–6692, Bangkok, Thailand. Association for Computational Linguistics.

Marc Fisher, John Woodrow Cox, and Peter Hermann. 2016. Pizzagate: From rumor, to hashtag, to gunfire in dc. *Washington Post*, 6:8410–8415.

Mohamed H. Gad-Elrab, Daria Stepanova, Jacopo Urbani, and Gerhard Weikum. 2019. Exfakt: A framework for explaining facts over knowledge graphs and text. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 87–95, New York, NY, USA. Association for Computing Machinery.

Boris Galitsky. 2025. 8 - truth-o-meter: Collaborating with llm in fighting its hallucinations. In William Lawless, Ranjeev Mittu, Donald Sofge, and Hesham Fouad, editors, *Interdependent Human-Machine Teams*, pages 175–210. Academic Press.

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

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

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 (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Vipin Gupta, Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. 2022. MMM: An emotion and novelty-aware approach for multilingual multimodal misinformation detection. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 464–477, Online only. Association for Computational Linguistics.

David Güera and Edward J. Delp. 2018. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. Lora+: Efficient low rank adaptation of large models. In *International Conference on Machine Learning*, pages 17783–17806. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR 2022*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. 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, Online. Association for Computational Linguistics.

Sujit Kumar, Anshul Sharma, Siddharth Hemant Khincha, Gargi Shroff, Sanasam Ranbir Singh, and Rahul Mishra. 2025. Sciclaimhunt: A large dataset for evidence-based scientific claim verification. *arXiv preprint arXiv:2502.10003*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Paul Mozur. 2018. A genocide incited on facebook, with posts from myanmar's military. *The New York Times*, 15(10):2018.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. 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, Marseille, France. European Language Resources Association.

Dan S. Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3141–3153, New York, NY, USA. Association for Computing Machinery.

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 (Volume 2: Short Papers)*, pages 476–483, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jasabanta Patro and Sabyasachee Baruah. 2021. A simple three-step approach for the automatic detection of exaggerated statements in health science news. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3293–3305, Online. Association for Computational Linguistics.

Jasabanta Patro and Pushpendra Singh Rathore. 2020. A sociolinguistic route to the characterization and detection of the credibility of events on twitter. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 241–250, New York, NY, USA. Association for Computing Machinery.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent convolutional strategies for face manipulation detection in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

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*, pages 624–643, Online. Association for Computational Linguistics.

Gautam Kishore Shahi and Durgesh Nandini. 2020. Fakecovid- A multilingual cross-domain fact check news dataset for COVID-19. In *Workshop Proceedings of the 14th International AAAI Conference on*

Web and Social Media, ICWSM 2020 Workshops, Atlanta, Georgia, USA [virtual], June 8, 2020.

Kai Shu, Ahmadreza Mosallanezhad, and Huan Liu. 2022. Cross-domain fake news detection on social media: A context-aware adversarial approach. In *Frontiers in fake media generation and detection*, pages 215–232. Springer.

Aryan Singhal, Thomas Law, Coby Kassner, Ayushman Gupta, Evan Duan, Aviral Damle, and Ryan Luo Li. 2024. Multilingual fact-checking using LLMs. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 13–31, Miami, Florida, USA. Association for Computational Linguistics.

Shivangi Singhal, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Factdrill: A data repository of fact-checked social media content to study fake news incidents in india. *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):1322–1331.

S. Suryavardan, Shreyash Mishra, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Naresh Reganti, Aman Chadha, Amitava Das, Amit P. Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Factify 2: A multimodal fake news and satire news dataset. In *Proceedings of De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2023, Washington DC, USA, February 14, 2023*, volume 3555 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Shivani Tufchi, Ashima Yadav, and Tanveer Ahmed. 2023. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval*, 12(2):28.

Juraj Vladika and Florian Matthes. 2023. Scientific fact-checking: A survey of resources and approaches. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6215–6230, Toronto, Canada. Association for Computational Linguistics.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024a. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, pages 2452–2463.

Xinyu Wang, Wenbo Zhang, and Sarah Rajtmajer. 2024b. Monolingual and multilingual misinformation detection for low-resource languages: A comprehensive survey. *arXiv preprint arXiv:2410.18390*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Deressa Wodajo and Solomon Atnafu. 2021. Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*.

Dustin Wright and Isabelle Augenstein. 2020. Claim check-worthiness detection as positive unlabelled learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022a. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. *arXiv preprint arXiv:2209.14642*.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022b. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the*

46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10331–10343.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. Metaadapt: Domain adaptive few-shot misinformation detection via meta learning. *arXiv preprint arXiv:2305.12692*.

Fengzhu Zeng and Wei Gao. 2023. Prompt to be consistent is better than self-consistent? few-shot and zero-shot fact verification with pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4555–4569, Toronto, Canada. Association for Computational Linguistics.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tong Zhang, Di Wang, Huanhuan Chen, Zhiwei Zeng, Wei Guo, Chunyan Miao, and Lizhen Cui. 2020. Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In *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 996–1011.

Xiaofan Zheng, Zinan Zeng, Heng Wang, Yuyang Bai, Yuhan Liu, and Minnan Luo. 2025. From predictions to analyses: Rationale-augmented fake news detection with large vision-language models. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 5364–5375, New York, NY, USA. Association for Computing Machinery.

Appendix

A Related works:

The NLP community has long been focusing on the core and peripheral tasks of automated fact-checking. Several datasets and methodologies have been proposed in the past, spanning different domains, languages, and task frameworks. In the following, we have reported the past literature for (i) monolingual fact-checking, (ii) multi-lingual and multi-modal fact-checking, (iii) explainable fact-checking, and (iv) associated tasks in this area. Apart from that, we have also reported how our work fills the current research gap.

A.1 Monolingual fact-checking:

Here, we reported the past works for monolingual fact-checking. Specifically, we focused on two aspects, i.e. datasets and approaches. We found several datasets focusing on text-based fact-checking in the monolingual domain. Some of the popular ones are, FEVER (Thorne et al., 2018), FEVEROUS (Aly et al., 2021), VITAMIN-C (Schuster et al., 2021), LIAR-RAW and RAW-FC (Yang et al., 2022b). Thorne et al. (2018) produced one of the earlier famous datasets, FEVER, consisting of around 185K, claims which were extracted from the Wikipedia corpus. Annotators tagged them with three distinct labels: 'Supported', 'Refuted' or 'NotEnoughInfo' (NEI). LIAR-RAW and RAWFC datasets were constructed by Yang et al. (2022b). RAW-FC consists of around 2K claims with associated raw reports collected from the Snopes [2] website. They were annotated for three labels, namely, 'true', 'half-true', and 'false'. On the other hand, the LIAR-RAW dataset was an extension of the LIAR-PLUS (Alhindi et al., 2018) dataset. The authors accompanied each claim with raw reports, and expert annotators labelled the claims with one label out of six: 'pants-fire', 'false', 'barely-true', 'half-true', 'mostly-true' and 'true'. The samples are domain-agnostic and mainly used to train general-domain fact-checking models. Additionally, we have domain-specific datasets as well. For example, PUBHEALTH dataset proposed by Kotonya and Toni (2020b) does fact-checking in the healthcare domain. It consists of 11.8K claims, where each claim is annotated and tagged with one of the following four labels: 'true', 'false', 'mixture', and 'unproven'. Similarly, the SciClaimHunt dataset by Kumar et al. (2025) focused on scientific claim verification. Here, each claim is labelled as positive or negative based on the evidence provided in the scientific research papers. A complete list of datasets focusing on automated fact-checking is reported in Guo et al. (2022); Vladika and Matthes

[2] https://www.snopes.com/

(2023); Zeng et al. (2021).

From a methodological point of view, the choice of fact-checking systems was closely tied to the datasets and task frameworks. Early approaches (DAGAN et al., 2010; Bowman et al., 2015) assessed whether each retrieved evidence supports or refutes a claim as an entailment task.

A.2 Mutli-lingual and multi-modal fact checking

The need to address misinformation across various languages and media formats have expanded the field of fact-checking into multilingual and multi-modal domains. In multilingual settings, datasets such as FakeCovid (Shahi and Nandini, 2020), which focuses on COVID-19 misinformation covering 40 languages, and X-Fact (Gupta and Srikumar, 2021) covering 25 languages and diverse domains. Prior surveys (Wang et al., 2024b; Singhal et al., 2024) provided a complete list of datasets and methods ranging from machine learning classifiers to LLM based techniques used for multi-lingual fact-checking. In parallel, a comprehensive survey by Akhtar et al. (2023a) highlighted the evolution of multi-modal fact checking domain. Several datasets have emerged to support this line of research such as FactDrill (Singhal et al., 2022) which combines video, audio, image, text, and metadata, while r/Fakeddit (Nakamura et al., 2020), MuMiN (Nielsen and McConville, 2022), MOCHEG dataset (Yao et al., 2023), and Factify-2 (Suryavardan et al., 2023) focused on image-text pairs only. Gupta et al. (2022) introduced a novel multi-lingual multimodal misinformation (MMM) dataset which integrated three Indian languages into multimodal fact-checking domain. In terms of modeling, early approaches used CNN-based models, such as ResNet (Sabir et al., 2019), and VGG16 (Amerini et al., 2019). Later on, studies integrated recurrent networks such as LSTM (Güera and Delp, 2018) to give better verdict. In the following years, researchers used CNN-LSTM hybrid (Tufchi et al., 2023) and transformers based models such as ViT (Wodajo and Atnafu, 2021). Recently, pretrained models have gained pace into multimodal fact checking (Zhang et al., 2020; Cekinel et al., 2025). Several studies (Akhtar et al., 2023b; Tufchi et al., 2023) highlighted the list of datasets and methods used for multi-modal fact checking.

A.3 Explainable fact-checking:

Explainability has emerged as a critical requirement for trustworthy fact-checking systems. Kotonya and Toni (2020a); Eldifrawi et al. (2024) illustrated the recent advancements in this domain through comprehensive surveys. Alhindi et al. (2018) created the LIAR-PLUS dataset, extending the LIAR dataset, by including justification from long ruling comments. Except for general domain, explanability has also entered into more intricate domains like healthcare (Kotonya and Toni, 2020b). In multi-modal contexts, MOCHEG (Yao et al., 2023) introduced the first end-to-end dataset to include structured explanations, bridging the gap between multi-modal fact verification and interpretability. Early methods relied on attention mechanisms (i.e., highlighting high-attention tokens) to pinpoint evidence span for explanation (Popat et al., 2018). But critiques said attention weights often misrepresent true model reasoning and lack accessibility for non-experts (Pruthi et al., 2020). Rule-based systems (Gad-Elrab et al., 2019) improved transparency but restricted scope to structured claims and pre-defined data. More recent textual explanation models face challenges like hallucination in abstractive outputs (Maynez et al., 2020). Out of them, our proposed method utilized the 'Justification-Then-Veracity' pipeline, as mentioned in Eldifrawi et al. (2024), for reliable fact-checking.

A.4 Associated tasks:

There are various associated tasks along with fact-checking such as claim check-worthiness, rumour detection, stance detection, etc. Claim check-worthiness (Wright and Augenstein, 2020) is required before jumping directly into fact verification. Another related task, rumour detection (Dougrez-Lewis et al., 2022; Gorrell et al., 2019) identifies unverified claims in real-time, often using propagation patterns. Stance detection gauges public reactions to prioritize claims (Baly et al., 2018). Exaggeration detection (Patro and Baruah, 2021) is another task where similar statements are compared to find which of them is a more exaggerated version. Credibility detection (Patro and Rathore, 2020) assesses the trustworthiness of topics discussed in social media. Clickbait detection (Chakraborty et al., 2016) identifies the online content (headlines or titles) specially designed to attract users to click.

A.5 Research gap:

Prior work by Yue et al. (2024) took retrieved evidences to generate supporting and opposing arguments for independent evaluation, and utilized few-shot prompting for claim verification. Similarly, Wang et al. (2024a) split relevant evidences into supporting or refuting categories relying on a complex attention mechanism. However, both of them suffer from key limitations: Yue et al. (2024)'s method did not consider dividing evidence set that may contain contradictory statements, whereas Wang et al. (2024a)'s complex attention mechanism discarded some useful information by focusing only on the top-k evidences based on the attention score. Here, we argue that VLLMs, with their broad understanding of language can instead analyse how each evidence supports (entails) or refutes the claim. Further, with their ability to process long input text, VLLMs can consolidate the whole support and refute evidence set to generate respective justification. To test this hypothesis, we compared it with various training-based and inference-based methods. To the best of our knowledge, our work is the first to use claim-evidence entailment in VLLMs for this task, ensuring no evidence is overlooked.

B Additional details on datasets:

In this section, we reported a detailed discussion of the considered datasets. Additionally, we have also presented some samples and statistics for the datasets for better illustration.

B.1 LIAR-RAW (Yang et al., 2022b):

LIAR-RAW (Yang et al., 2022b) consists of 12,590 claims, each paired with raw reports collected from news articles, press releases, and web pages. LIAR-RAW builds upon the LIAR-Plus dataset (Alhindi et al., 2018) by adding crowd-sourced raw reports. Authors have retrieved up to 30 raw reports for each claim via the Google API using claim keywords. Claims were collected from a well-known fact-checking website, i.e. *Politifact*[3], which provided gold veracity labels. To ensure quality, they excluded fact-checking site reports, those published after the verdict, and reports under 5 words or over 3,000 words. After filtering, the final dataset had 10,065 training instances, 1,274 validation instances, and 1,251 test instances. Expert annotators manually assigned one of six fine-grained veracity labels to each claim. The labels they considered are: **"pants-fire"** (completely false), **"false"**, **"barely-true"** (contains minimal truth but is mostly false), **"half-true"** (equally true and false, with a significant mix of both), **"mostly-true"** (predominantly true with minor inaccuracies) and **"true"**. A sample from the dataset is presented in Table 6 for illustration. We have also studied the claim-evidence distribution, where we clubbed the claims based on the number of evidence sentences they have. The distribution is illustrated in Figure 2. We found that (i) the majority of claims have no evidence sentences, (ii) more than a thousand of claims have one evidence sentence, and (iii) 170+ claims have more than fifty evidence sentences associated with them.

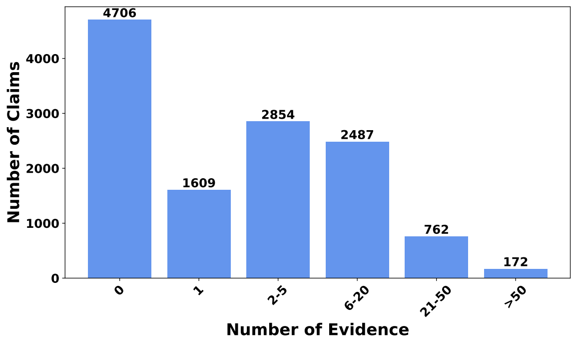


Figure 2: Bar plot showing number of claims v/s number of evidences for LIAR-RAW dataset

B.2 RAW-FC (Yang et al., 2022b):

RAWFC dataset, introduced by (Yang et al., 2022b), has 2,012 claims. Authors collected claims from *Snopes*[4] website and retrieved relevant raw reports using claim keywords via the Google API. For each claim, up to 30 raw reports were gathered from various web sources. To ensure quality, reports from fact-checking sites and those published after the fact-checking verdict were excluded. Reports shorter than 5 words or longer than 3,000 words were also removed. Expert annotators manually assigned one of the three veracity labels to each claim. The labels used for the annotation are: **"true"** (entirely accurate), **"half-true"** (partially true but includes misleading information), and **"false"** (entirely false). A sample from the dataset is presented in Table 7 for illustration. We have studied the claim-evidence distribution for this dataset as well. We created buckets based on evidence sentence counts and gathered the claims

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

[4] https://www.snopes.com

Dataset: LIAR-RAW (Yang et al., 2022b)

Event_id: “5209.json”

Claim: “Suzanne Bonamici supports a plan that will cut choice for Medicare Advantage seniors.”

Label: “half-true”

Explain: “The Affordable Care Act was designed to save money by slowing future spending, including future spending on Medicare Advantage plans. But spending still goes up. In addition, many outside factors can affect the cost and range of benefits, making it impossible to know how Medicare Advantage might change. While the statement from Cornilles is partially accurate, it is taken out of context and ignores important details on a politically volatile subject.”

Evidences:
E1: “So I hope , a we move forward , we will focus on that approach to modernize Medicare and Medicaid , an approach that improve the quality of care while we reduce the cost of care , rather than simply offload those cost onto senior .”,
E2: “ So some of the save we achieve , a significant amount of save we achieve , be in reduce these overpayment , these huge subsidy , to the private Medicare Advantage plan .”,
E3: “In addition to the issue with Part D benefit design and plan flexibility , there be transaction such a rebate , pharmacy fee , and other form of compensation that occur in the supply chain that pose several issue .”,
E4: “ The ACA have help slow the growth in health care cost , it be close the doughnut hole for senior , and have encourage and improved access to mental health service and preventive care .”

Table 6: Example entry from the LIAR-RAW dataset. It contains five key fields: (i) an identifier ‘*Event_id*’, (ii) the ‘*claim*’ to be fact-checked, (iii) the ground-truth ‘*Label*’, (iv) an explanation (‘*Explain*’) that clarifies its ground truth, and (v) the evidence sentences (*E1–E4*) extracted from the underlying fact-checking article.

belonging to individual buckets. The distribution is illustrated in Figure 3. We found that (i) around 750 claims have twenty to fifty evidence sentences associated with them, and (ii) more than five hundred claims have fifty or more evidence sentences associated with them.

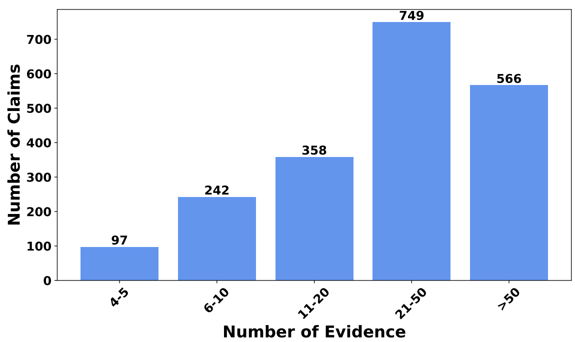


Figure 3: Bar plot showing number of claims v/s number of evidences for RAW-FC dataset

C Additional Details on Experiments:

In this section, we have reported additional details on our experiments. Particularly, we described individual **IBEs**, provided additional details on baselines and experimental setups.

C.1 IBE-1: Zero-shot prompting:

In this experiment, we have tried to answer *"How well the VLLMs can predict the claim veracity if raw evidence sentences are provided in the input?"* To conduct this experiment, we prompted five VLLMs by giving claims and associated evidence sentences as input. Some of the prompt samples are showcased in Figure 4 and Figure 5 for illustration. In past, researchers have tried zero-shot prompting for this task; however, either (i) they gave only claims without evidence sentences (Zhang and Gao, 2023), or (ii) they relied on evidence retrieved from web searches (Cheung and Lam, 2023).

Dataset: RAW-FC (Yang et al., 2022b)

Event_id: "247609"

Claim: "Right-wing commentator Bill Mitchell tweeted \"someone let me know\" when 55,000 Americans have died from COVID-19 coronavirus disease."

Label: "true"

Explain: "In late April 2020, a month-old tweet posted by right-wing Twitter commentator Bill Mitchell received new attention amid tragic circumstances.\n Feeding on a conservative media talking point that the COVID-19 coronavirus disease pandemic isn\u2019t as serious as officials say, Mitchell tweeted on March 21, 2020:\u201cWhen COVID-19 reaches 51 million infected in the US and kills 55,000, someone let me know.\u201d\nSome readers asked if the statement (displayed above) was a real tweet posted by Mitchell, and it is.\nSadly, as of this writing, the U.S. has surpassed 55,000 fatalities from COVID-19. The Centers for Disease Control and Prevention (CDC) reports that as of May 1, 2020, there are 1,062,446 COVID-19 cases in the U.S. and 62,406 deaths.\nSocial media users wasted no time posting comments directed at Mitchell, reminding him of his statement.\n\nBecause the tweet in question was published from Mitchell\u2019s Twitter account, we rate this claim\u201cCorrect Attribution.\u201d"

Evidences:

E1: "On March 21, Bill Mitchell tweet, \u2018 When COVID-19\u2026kills 55,000 , let me know.",

E2: " In it , Hannan write that \u2018 [COVID-19] be unlikely to be as lethal a the more common form of influenza that we take for granted.",

E3: " While I \u2019 d normally be content to mock conspiracy theorist \u2014 I set up a Twitter account to make fun of bad COVID-19 take \u2014 spread false information about the pandemic be dangerous , and merit rebuttal.",

E4: " \u201d Owens delete the tweet after a Twitter user observe that her claim be not only false , but appear to be base on a lazy misreading of a Google search result .",

E5: " Given the dire situation , it seem worth know if the severity of the pandemic finally have penetrate the bubble of the most extreme coronavirus scoffer ; if these people and their follower be ignore safety measure because they believe the pandemic be a false flag to take away their gun or a Deep State plot to take down Trump , then they risk contribute to the disease \u2019 s spread ."

Table 7: Example entry from the RAW-FC dataset. It contains five key fields: (i) an identifier '*Event_id*', (ii) the '*claim*' to be fact-checked, (iii) the ground-truth '*Label*', (iv) an explanation ('*Explain*') that clarifies its ground truth, and (v) the evidence sentences (*E1–E4*) extracted from the underlying fact-checking article.

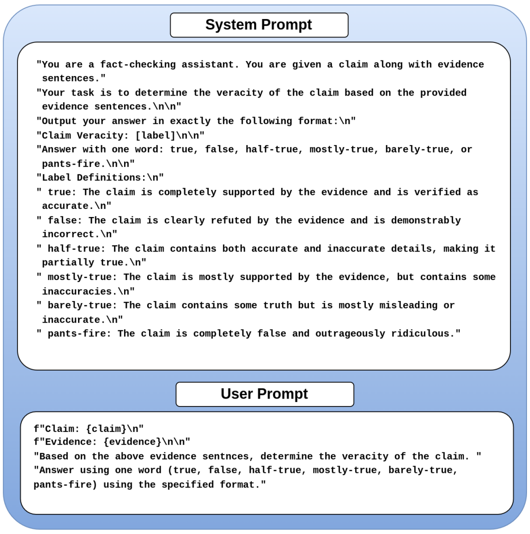


Figure 4: IBE-1 (LIAR-RAW): Prompt used to predict the veracity of the claim using zero-shot prompting based on the given evidence sentences and the claim.

C.2 IBE-2: Zero-shot prompting with overall understanding:

In this experiment, we have tried to answer *"Can zero-shot prompting improve the veracity prediction performance if claim-evidence relation understandings generated by VLLMs are given as input?"*. To conduct this experiment, we followed a two-step prompting strategy. First, we prompted the VLLMs to generate the claim-evidence understanding like we did in **TBE-2**. In the next step, we prompted them again with the claim and the generated understanding to predict the veracity labels. Some of the prompt samples are presented in the Figure 7 and Figure 8. To the best of our knowledge, nobody has attempted this in past.

C.3 IBE-3: CoT (Wei et al., 2022) prompting with overall understanding:

In this experiment, we tried to answer *Can CoT-based prompting improve the veracity prediction performance if claim-evidence understanding is given in the input?*. Here we followed similar steps as mentioned in **IBE-2**, except in the prompt, we asked the VLLMs to generate step-by-step reasoning behind their predictions. Prior work (Zhang and Gao, 2023) did a similar attempt. However, they gave evidence collected from web searches. We, on the other hand, relied on the claim-evidence understanding generated by VLLMs. Some of the prompt samples are presented in Figure 10 and Figure 11 for illustration.

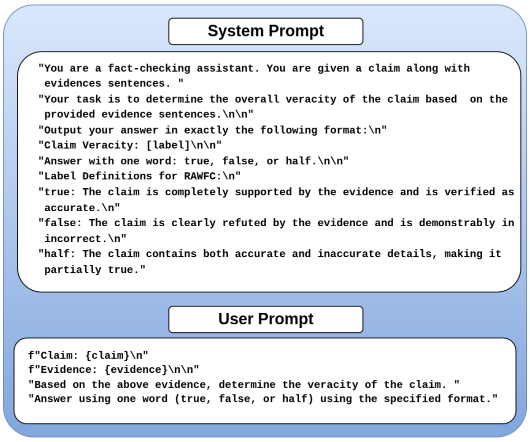


Figure 5: IBE-1 (RAW-FC): Prompt used to predict the veracity of the claim using zero-shot prompting based on the given evidence sentences and the claim.

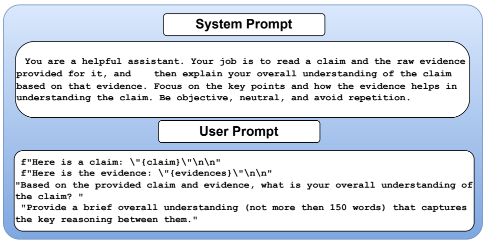


Figure 6: IBE-2: First step, prompt used to generate an overall understanding of the claim based on the provided evidence sentences.

C.4 IBE-4: Prompting based on entailment:

In the last inference-based experiment, we tried to answer *"Can prompting with VLLM generated entailed-justifications enhance language models' ability to predict claim veracity?"* To conduct this experiment, we have followed a three-step approach similar to **TBE-3** as described in Section 3.1.3. While the initial two steps, i.e. (i) to classify the evidence sentences as supporting or refuting a given claim and (ii) generating the justifications based on classified evidence sentences, are exactly the same, in the last step, instead of training, we prompted the VLLMs to generate the veracity. In past, researchers have prompted the VLLMs to find entailment for tasks like claim matching (Choi and Ferrara, 2024b) and counterfactual generation (Dai et al., 2022). However, to the best of our knowledge, we are the first to (i) apply it to classify each evidence into supporting or refuting categories, and (ii) generate entailed justifications and use them for fact verification. The detailed

Figure 7: IBE-2 (LIAR-RAW): Second step, prompt using zero-shot prompting to directly predict the claim's veracity based on the overall understanding.

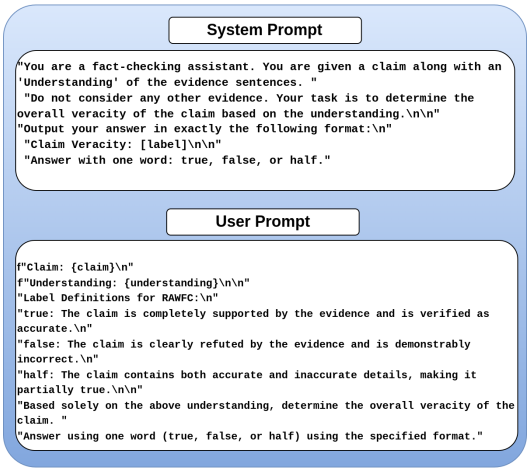


Figure 8: IBE-2 (RAW-FC): Second step, using zero-shot prompting to predict the claim's veracity from the overall understanding.

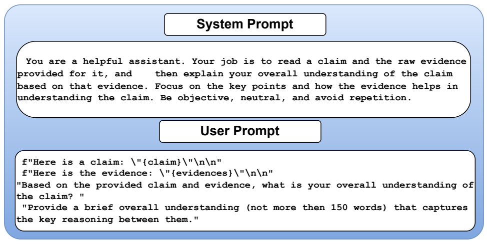


Figure 9: IBE-3: First step, prompt used to generate an overall understanding of the claim based on the provided evidence sentences.

approach is illustrated in Fig. 1 (c). Some of the prompt samples are show cased in the Figure 12, Figure 13, Figure 14, Figure 15 and 16.

C.5 Additional details on baselines:

In this section, we reported the details of individual baseline methods considered in our study.

- **HiSS**: It is proposed by Zhang and Gao (2023). Here, authors have proposed a hierarchical step-by-step (a.k.a. *HiSS*) method where they first decomposed a claim into smaller sub-claims using few-shot prompting. Then they verified each sub-claim step-by-step by raising and answering a series of questions. For each question, they prompted the language models to assess if it is confident in answering or not, and if not, they gave the question to a web search engine. The search results were then inserted back into the ongoing prompt to continue the verification process. Finally, using that information, language models predicted the veracity label for the whole claim.

- **FactLLaMa**: It is proposed by Cheung and Lam (2023). Their approach has two components: (i) generation of prompts (having instructions, claims and evidence pieces), and (ii) instruction-tuning of a generative pretrained language model. In the first component, to create the prompt samples, they combined the instruction, evidence, and input claim into a single sequence, with special tokens separating them. While the instruction guides how to incorporate the evidence for fact-checking, evidence contains relevant information retrieved from search engines, using the Google API. As a part of the second component they we fine-tuned the LoRA adapter with Llama as the language model.

- **RAFTS**: It stands for *retrieval augmented fact verification through the synthesis of contrastive arguments*. It is proposed by Yue et al. (2024), where authors retrieve relevant documents and perform a few-shot fact verification using pretrained language models. RAFTS have three components, (i) demonstration retrieval, where relevant examples were collected to included in the input contexts, (ii) document retrieval, where authors proposed a *retrieve and re-rank* pipeline (using RAG framework) to accurately identify

System Prompt

```
"You are a fact-checking assistant. You are given a claim along with an
'Understanding' of the evidence sentences for the claim. "
"Do not consider any other evidence. Your task is to determine the  veracity of
the claim based on the understanding.\n\n"
"First, explain your reasoning step-by-step using only the understanding
provided.\n"

"Then, determine the claim's veracity using one of the following six
categories:\n\n"
"1) pants-fire  : The claim is ridiculously or outrageously untrue.\n"
"2) false       : The claim is entirely incorrect and unsupported by
                        evidence.\n"
"3) barely-true : The claim has a small element of truth but is mostly
                          misleading.\n"
"4) half-true   : The claim mixes true and false information in significant
                         amounts.\n"
"5) mostly-true : The claim is generally correct but contains minor
                          inaccuracies.\n"
"6) true        : The claim is completely accurate and fully supported by
                       evidence.\n\n"
 "Finally, output your answer in exactly the following format:\n"
 "Claim Veracity: [label]\n\n"
 "where [label] must be one of (pants-fire, false, barely-true, half-true,
  mostly-true, or true)."
```

User Prompt

```
f"Claim: {claim}\n"
f"Understanding: {understanding}\n\n"
"First, explain your reasoning based only on the
understanding provided above.\n"
"Then, determine the overall veracity of the claim using one of the following
        labels:\n"
     "- pants-fire\n"
     "- false\n"
     "- barely-true\n"
     "- half-true\n"
     "- mostly-true\n"
     "- true\n\n"
     "Finally, answer in the following format:\n"
     "Claim Veracity: [label]"
```

Figure 10: IBE-3 (LIAR-RAW): Second step, prompt using Chain-of-Thought reasoning to predict the veracity of the claim from the overall understanding.

relevant documents for the input claim, and (iii) few-shot fact verification using supporting and opposing arguments derived from the facts within the collected documents. However, unlike in our work, they didn't rely on supporting and opposing justifications (a.k.a. entailed justifications) in their final steps.

- **L-Defense**: The LLM-equipped defense-based explainable fake news detection approach (L-Defense) was proposed by Wang et al. (2024a). Their approach shares some similarity with RAFTS at a philosophical level. The framework consists of three components: (i) a competing evidence extractor, (ii) a prompt-based reasoning module, and (iii) a defense-based inference module. In the competing evidence extractor, authors deployed a natural language inference (NLI) module to associate a "true" or "false" label to each claim for each associated evidence sentence. The NLI module gave two top-k evidence sentence sets, each stating a claim to be "true" or "false. In the reasoning module, they prompted language models to generate two separate explanations, each for stating a claim to be "true" and "false based on the respective top-k evidence set. Finally, in the defense-based inference module, they trained the transformer encoders by taking claims and associated two explanations to predict the veracity. Our approach is different from the L-Defense as we (i) deployed VLLMs to classify each evidence, (ii) we considered all evidences associated with a claim, and (iii) we fine-tuned LLMs and adapters with VLLMs for veracity prediction (in **TBE-3**).

System Prompt

```
"You are a fact-checking assistant. You are given a claim along with an
'Understanding' of the claim. "
"Do not consider any other evidence. Your task is to determine the overall
 veracity of the claim based on the understanding.\n\n"
"First, explain your reasoning in detail based only on the understanding.\n"
"Then output the claim veracity label in **exactly** the following format:\n"
"Claim Veracity: [label]\n\n"
"Answer with one word: true, false, or half."
```

User Prompt

```
  f"Claim: {claim}\n"
  f"Understanding: {understanding}\n\n"
 "Label Definitions:\n"
 "true: The claim is completely supported by the evidence and
  accurate.\n"
 "false: The claim is clearly refuted by the evidence and  incorrect.\n"
 "half: The claim contains both accurate and inaccurate details, making it
  partially true.\n\n"
 "First, explain your reasoning step-by-step using only the
  understanding above.\n"
  "Then, based on your reasoning, provide the final claim veracity in the
  format:\n"
  "Claim Veracity: [true/false/half]
```

Figure 11: IBE-3 (RAW-FC): Second step, prompt using Chain-of-Thought reasoning to predict the veracity of the claim from the overall understanding.

System Prompt

```
  You are an evidence classifier that determines whether a sentence supports or
refutes a given claim.
```

User Prompt

```
   Task: Analyze the sentence below and classify its relationship to the given
 claim.
  Classification Rules:
  1. SUPPORT - The sentence provides evidence that supports the claim.
  2. REFUTE- The sentence provides evidence that contradicts or refutes the
claim.

  Claim: {claim}
  Label for Claim: {label}
  Sentence to classify: {sent}

  Carefully analyze the sentence and respond with:
  1. A classification (SUPPORT or REFUTE).
```

Figure 12: IBE-4: First step prompt used to classify each evidence sentence as supporting or refuting the claim.

C.6 Experimental set-up:

C.6.1 Training/ test/ validation splits:

We have used the training, validation, and test splits originally provided by Yang et al. (2022b) in our experiments. They are essential for the comparison of our models with the baselines. The distribution of samples falling under each label and training splits are reported in Table 10.

C.6.2 Language models:

For prompting, we used five very large language models (VLLMs). We call them so because of

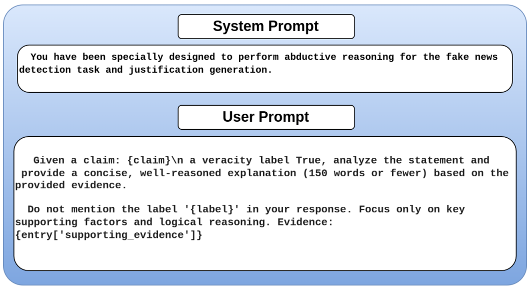


Figure 13: IBE-4, Second step prompt used to generate a supporting justification based on the supporting evidence sentences.

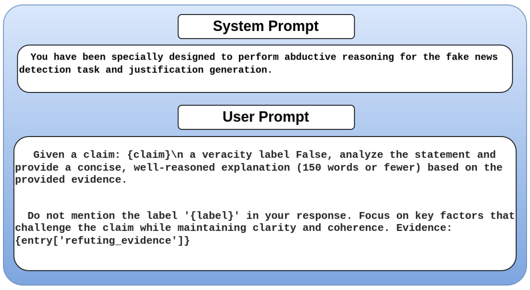


Figure 14: IBE-4, Second step prompt used to generate a refuting justification based on the refuting evidence sentences.

their immense parameter size, which prohibits us from fine-tuning them like we do for normal BERT-based language models. We used five VLLMs i.e. Mistral (Jiang et al., 2023), Llama (AI@Meta, 2024), Gemma (Team et al., 2024), Qwen (Yang et al., 2024) and Falcon (Almazrouei et al., 2023) in all of our experiments. For fine-tuning, we used LLMs RoBERTa and XLNet in **TBE-2** and **TBE-3**. Apart from that, we also used adapter-based (LoRA (Hu et al., 2022) and LoRA+ (Hayou et al., 2024) adapters) fine-tuning methods for the considered VLLMs. The details of LLMs and VLLMs such as their versions and maximum input size they can take are reported in Table 9.

C.6.3 Hyperparameter details:

We did an extensive hyperparameter search that led to the optimal performance of our models. The list of hyperparameters for which we trained our models is presented in Table 8. For the VLLMs, we kept the temperature constant at '0.001' for consistency. We conducted all our experiments on two NVIDIA A100 80GB GPU cards.

C.6.4 Evaluation metrics:

Since we are working in a multi-class classification framework for veracity prediction, we used standard metrics such as macro-precision (MP),

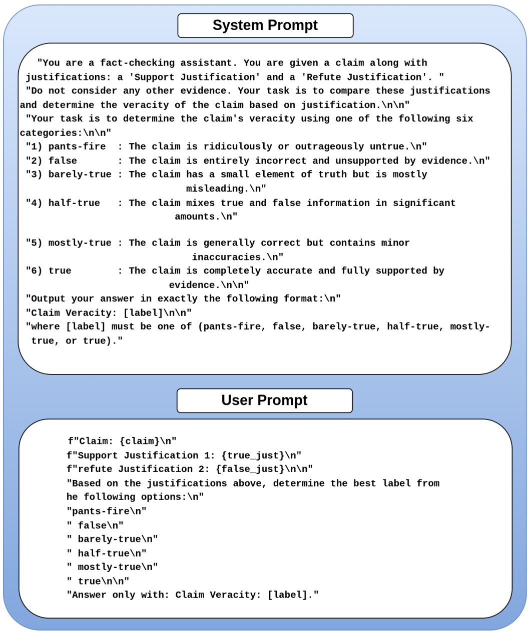


Figure 15: IBE-4 (LIAR-RAW): Third step prompt used to predict the veracity of the claim using the supporting and refuting justifications.

Parameters	**Values**
Learning rate	{2e-6, 2e-5, 1e-5}
Optimizer	AdamW, Adam
Batch size	8, 16
Patience (Early stop)	2, 3
lora_rank	8
Learning rate	{1e-5,1e-4}
lr_scheduler_type	cosine
bf16	true

Table 8: Hyperparameters explored during model training and evaluation.

macro-recall (MR), and macro-f1 (MF) to evaluate our models. To evaluate model explainability, we first concatenated the supporting and refuting entailed justifications (generated as part of TBE-3) and considered them as model explanations. The evaluation was done with two types of strategies: (i) checking lexical overlap and semantic matching, and (ii) doing subjective evaluation by VLLMs. To check lexical overlap, we used several standard evaluation metrics such as ROUGE-1 (R_1), ROUGE-2 (R_2), ROUGE-L (R_L) (Lin, 2004) and BLEU (Papineni et al., 2002). While R_1 and R_2 measure the overlap of unigrams and bigrams between predicted and gold explanations, R_L measures the longest common subsequence. BLEU, on the other hand, measures the precision of matching n-grams between predicted and gold explanations.

System Prompt

```
"You are a fact-checking assistant. You are given a claim along with two
 generated justifications: a 'Support Justification' and a 'Refute
 Justification'. "
"Do not consider any other evidence. Your task is to compare these
 justifications   and determine the overall veracity of the claim based
 justification.\n\n"
"Output your answer in exactly the following format:\n"
"Claim Veracity: [label]\n\n"
"Answer with one word: true, false, or half."
```

User Prompt

```
 f"Claim: {claim}\n"
 f"Support Justification: {support_just}\n"
 f"Refute Justification: {refute_just}\n\n"
 "Label Definitions for RAWFC:\n"
 "true: The claim is completely supported by the evidence and is verified as
accurate.\n"
 "false: The claim is clearly refuted by the evidence and is demonstrably
incorrect.\n"
 "half: The claim contains both accurate and inaccurate details, making it
partially true.\n\n"
 "Based on the above justifications, determine the overall veracity of the
claim. "
 "Answer using one word (true, false, or half) using the specified format."
```

Figure 16: IBE-4 (RAW-FC): Third step prompt used to predict the veracity of the claim using the supporting and refuting justifications.

Model	Max. length	Version
RoBERTa	1k	roberta-large
XLNet	1k	xlnet-large-cased
Mistral	32k	mistralai/Mistral-7B-Instruct-v0.3
Llama	128k	meta-llama/Llama-3.1-8B-Instruct
Gemma	8k	google/gemma-7b-it
Qwen	1M	Qwen/Qwen2.5-7B-Instruct-1M
Falcon	8k	tiiuae/Falcon3-7B-Instruct

Table 9: Language model details. Here, 'Max. length' denotes the maximum sequence length allowed by the particular model.

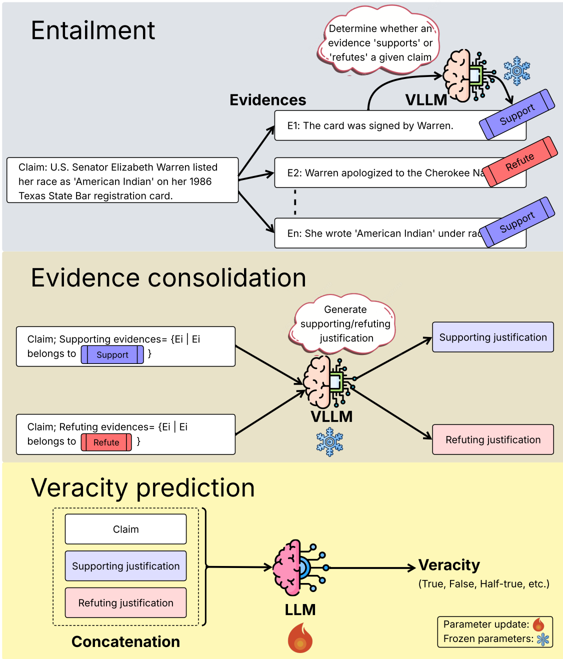


Figure 17: A schematic diagram explaining the TBE-3 pipeline which breaks down into three stages. (a) *Entailment*: Given a claim and its corresponding evidences, a VLLM first distinguishes evidences into supporting/ refuting evidence via entailment. (b) *Evidence consolidation*: Then the same VLLM consolidates these two groups into concise supporting and refuting justification. (c) *Veracity prediction*: Using the claim and both justifications, an LLM is trained to predict veracity.

To measure the semantic matching between predicted and gold explanations, we deployed BERT score (Zhang et al., 2019). It generates contextual embeddings of predicted and gold explanations and calculates cosine similarity between them. To do the subjective evaluation, we prompted the considered VLLMs (a.k.a evaluating VLLMs) to asses the model explanations generated by each VLLM (generating VLLM). The VLLMs were asked to asses across five dimensions (i) informativeness, (ii) logicality, (iii) objectivity, (iv) readability and (v) accuracy (Zheng et al., 2025). The prompt template used for the assessment is presented in Figure 25.

D Additional Results and Discussion:

D.1 Additional observations from TBE-1:

- We found that Mistral trained with LoRA adapter resulted in the highest macro-precision (**0.44**) for the LIAR-RAW dataset. While it surpasses the performance of baselines such as FactLLaMA (MP : **0.32**, **37.50**% $\uparrow$) and L-Defense (MP : **0.31**, $\sim$ **41.94**% $\uparrow$), it is still behind HiSS (MP : **0.46**) and RAFTS (MP : **0.47**). For RAWFC dataset, however, many models, such as Mistral trained with LoRA adapter (MP : **0.69**, $\sim$ **11.29**% $\uparrow$), Llama trained with LoRA (MP : **0.68**, $\sim$ **9.68**% $\uparrow$) and LoRA+ adapters (MP : **0.67**, $\sim$ **8.06**% $\uparrow$), Qwen trained with LoRA (MP : **0.67**, $\sim$ **8.06**% $\uparrow$) and LoRA+ (MP : **0.70**, $\sim$ **12.90**% $\uparrow$) adapters, outperformed the best performance reported by the baselines (MP : **0.62**). Qwen trained with LoRA+ adapter achieved the highest overall performance (MP : **0.70**).

- We observed that Llama trained with LoRA adapter resulted in the highest macro-recall (**0.30**) for the LIAR-RAW dataset. It falls behind the macro-recall reported by all baselines HiSS (MR : **0.31**), FactLLaMa (MR : **0.32**), L-Defense (MR : **0.31**), and RAFTS (MR : **0.37**). However, for the RAW-FC dataset, we observed that many models, such as Mis-

Dataset	Label	Train	Val	Test
	T	1647	169	205
	MT	1950	251	238
LAIR-RAW	HT	2087	244	263
(Yang et al., 2022b)	BT	1611	236	210
	F	1958	259	249
	PF	812	115	86
	T	561	67	67
RAW-FC	HT	537	67	67
(Yang et al., 2022b)	F	514	66	66

Table 10: Train, val and test distributions. Notations: **T** for True, **MT** for Mostly-true, **HT** for Half-true, **BT** for Barely-true, **F** for False, **PF** for Pants-fire and **T** for True.

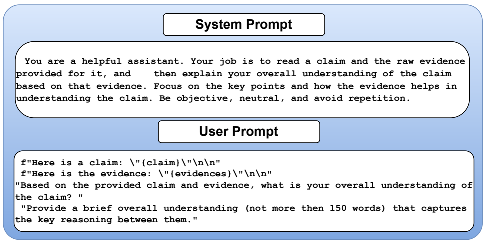


Figure 18: TBE-2: Prompt for generating overall understanding from the claim and evidence sentences.

tral trained with LoRA adapter (MR : **0.65**, $\sim$ **6.56**$\%$ $\uparrow$), Llama trained with LoRA (MR : **0.64**, $\sim$ **4.92**$\%$ $\uparrow$) and LoRA+ adapters (MR : **0.64**, $\sim$ **4.92**$\%$ $\uparrow$), Qwen trained with LoRA adapter (MR : **0.66**, $\sim$ **8.20**$\%$ $\uparrow$) outperformed the best reported macro-recall score by the baselines (**0.61**). Qwen trained with LoRA adapter achieved the highest overall performance (MR : **0.66**).

D.2 Additional observations from TBE-2:

- We found that Mistral trained (with Mistral understandings) with the LoRA adapter resulted in the highest macro-precision (**0.36**) for the LIAR-RAW dataset. While it outperforms the baselines FactLLaMA (MP : **0.32**) and L-Defense (MP : **0.31**) models, it is still behind HiSS (MP : **0.46**) and RAFTS (MP : **0.47**). However, for the RAW-FC dataset, we observed that many models, such as XLNet fine-tuned with Llama understandings (MP : **0.63**, $\sim$ **1.61**$\%$ $\uparrow$), Mistral trained (with Mistral understandings) with LoRA+ (MP : **0.66**, $\sim$ **6.45**$\%$ $\uparrow$) and Llama trained (with Llama understandings) with LoRA+ (MP : **0.73**, $\sim$ **17.74**$\%$ $\uparrow$) adapters outperformed the best reported macro-precision by the baselines (MP : **0.62**). Llama trained (with Llama understandings) with LoRA+ adapter resulted in the overall highest macro-precision (MP : **0.73**).

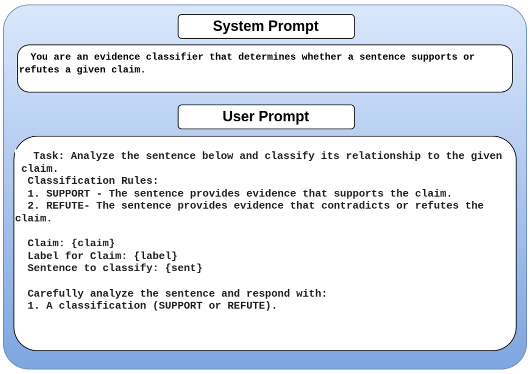


Figure 19: TBE-3: First step prompt used to classify each evidence sentence as supporting or refuting the claim.

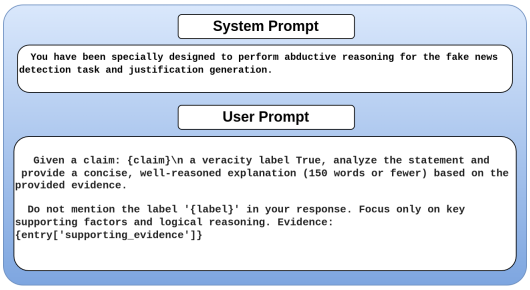


Figure 20: TBE-3: Second step prompt used to generate a supporting justification based on the supporting evidence sentences.

- We found that Mistral trained (with Mistral understandings) with LoRA adapter gave us the highest macro-recall (**0.32**) for LIAR-RAW dataset. While this performance is comparable to the performance of baselines HiSS (MR : **0.31**), FactLLaMA (MR : **0.32**), and L-Defense (MR : **0.32**), it is still behind RAFTS (MR : **0.37**). However, for the RAW-FC dataset, we observed that many models, such as Llama trained (with Llama understandings) with LoRA+ adapter (MR : **0.71**, $\sim$ **16.39**$\%$ $\uparrow$) and XLNet fine-tuned with Llama understandings (MR : **0.63**, $\sim$ **3.27**$\%$ $\uparrow$) outperformed the best reported macro-recall score by the baselines (MR : **0.61**). Llama trained (with Llama understandings) with LoRA+ adapter achieved the highest macro-recall overall (MR : **0.71** $\sim$ **16.39**$\%$ $\uparrow$).

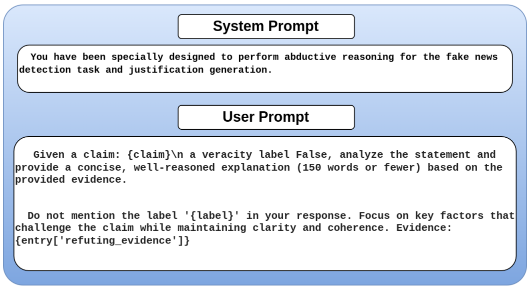


Figure 21: TBE-3: Second step prompt used to generate a refuting justification based on the refuting evidence sentences.

D.3 Additional observations from TBE-3:

- We found that XLNet fine-tuned with Llama entailed justifications gave the highest macro-precision (**0.55**) for the LIAR-RAW dataset. Many models, such as RoBERTa fine-tuned with Mistral (MP : **0.48**, $\sim$ **2.12**% $\uparrow$), Llama (MP : **0.53**, $\sim$ **12.76**% $\uparrow$), Gemma (MP : **0.49**, $\sim$ **4.25**% $\uparrow$), Qwen (MP : **0.48**, $\sim$ **2.12**% $\uparrow$), and Falcon (MP : **0.49**, $\sim$ **4.25**% $\uparrow$) entailed justifications, XLNet fine-tuned with Mistral (MP : **0.49**, $\sim$ **4.25**% $\uparrow$), Llama (MP : **0.55**, $\sim$ **17.02**% $\uparrow$), and Qwen (MP : **0.50**, $\sim$ **6.38**% $\uparrow$) entailed justifications and Llama trained (with Llama entailed justifications) with LoRA+ adapter (MP : **0.49**, $\sim$ **4.25**% $\uparrow$) surpassed the best reported macro-precision score of baselines (MP : **0.47**). Similarly, for the RAW-FC dataset, we observed that many models, such as such as RoBERTa fine-tuned with Mistral (MP : **0.83**, $\sim$ **33.87**% $\uparrow$), Llama (MP : **0.88**, $\sim$ **41.93**% $\uparrow$), Qwen (MP : **0.73**, $\sim$ **17.74**% $\uparrow$), and Falcon (MP : **0.65**, $\sim$ **4.83**% $\uparrow$) entailed justifications, XLNet fine-tuned with Mistral (MP : **0.83**, $\sim$ **33.87**% $\uparrow$), Llama (MP : **0.88**, $\sim$ **41.93**% $\uparrow$), Qwen (MP : **0.70**, $\sim$ **12.90**% $\uparrow$), and Falcon (MP : **0.76**, $\sim$ **22.58**% $\uparrow$) entailed justifications, Mistral trained (with Mistral entailed justifications) with LoRA (MP : **0.69**, $\sim$ **11.29**% $\uparrow$), Llama trained (with Llama entailed justifications) with LoRA (MP : **0.63**, $\sim$ **1.61**% $\uparrow$) and LoRA+ adapters (MP : **0.84**, $\sim$ **35.48**% $\uparrow$), respectively, surpassed the best macro-precision reported by the baselines (MP : **0.61**). RoBERTa and XLNet fine-tuned with Llama entailed justifications gave the highest macro-precision (MP : **0.88**).
- We observed that XLNet fine-tuned with Llama entailed justifications gave the highest macro-recall (**0.54**) for the LIAR-RAW dataset. Many models, such as RoBERTa fine-tuned with Mistral (MR : **0.47**, $\sim$ **27.02**% $\uparrow$), Llama (MR : **0.53**, $\sim$ **43.24**% $\uparrow$), Gemma (MR : **0.50**, $\sim$ **35.14**% $\uparrow$), Qwen (MR : **0.47**, $\sim$ **27.02**% $\uparrow$), and Falcon (MR : **0.43**, $\sim$ **16.22**% $\uparrow$) entailed justification, XLNet fine-tuned with Mistral (MR : **0.47**, $\sim$ **27.02**% $\uparrow$), Llama (MR : **0.54**, $\sim$ **45.95**% $\uparrow$), Gemma (MR : **0.43**, $\sim$ **16.22**% $\uparrow$), Qwen (MR : **0.48**, $\sim$ **29.73**% $\uparrow$), and Falcon (MR : **0.44**, $\sim$ **18.92**% $\uparrow$) entailed justifications, Llama trained (with Llama entailed justifications) with LoRA+ adapter (MR : **0.50**, $\sim$ **35.14**% $\uparrow$) and Falcon trained (with Falcon entailed justifications) with LoRA+ adapter (MR : **0.39**, $\sim$ **5.41**% $\uparrow$) surpassed the best reported macro-recall score of the baselines (MR : **0.37**). Similarly, for the RAW-FC dataset, we observed that many models, such as such as RoBERTa fine-tuned with Mistral (MR : **0.82**, $\sim$ **34.43**% $\uparrow$), Llama (MR : **0.88**, $\sim$ **44.26**% $\uparrow$), Qwen (MR : **0.71**, $\sim$ **16.39**% $\uparrow$), and Falcon (MR : **0.64**, $\sim$ **4.92**% $\uparrow$) entailed justifications, XLNet fine-tuned with Mistral (MR : **0.82**, $\sim$ **34.43**% $\uparrow$), Llama (MR : **0.88**, $\sim$ **44.26**% $\uparrow$), Qwen (MR : **0.70**, $\sim$ **14.75**% $\uparrow$), and Falcon (MR : **0.75**, $\sim$ **22.95**% $\uparrow$) entailed justifications, Llama trained (with Llama entailed justifications) with LoRA (MR : **0.62**, $\sim$ **1.64**% $\uparrow$) and LoRA+ adapter (MR : **0.82**, $\sim$ **34.43**% $\uparrow$) surpassed the best reported macro-precision score of baselines (MR : **0.61**). RoBERTa and XLNet fine-tuned on Llama entailed justifications gave the highest overall performance (MR : **0.88**).

D.4 Additional observations from IBEs:

- For LIAR-RAW dataset, we observed that Qwen gave the highest macro-precision (**0.27**) in IBE-1. While in IBE-2, Llama achieved the highest macro-precision (MP : **0.27**), Mistral gave the highest macro-precision in IBE-3 (MP : **0.27**) and IBE-4 (MP : **0.27**). However, none of them could surpass the baselines. In contrast, for RAW-FC dataset, Qwen (MP : **0.61**), Llama (MP : **0.62**), and Falcon (MP : **0.60**) achieved highest macro-precision in IBE-1, IBE-2, and IBE-3 respectively. They are comparable to the highest macro-precision reported by the baseline (MP : **0.62**). Similarly, Qwen (MP : **0.50**) got the best macro-precision in IBE-4, which is far behind the baselines.

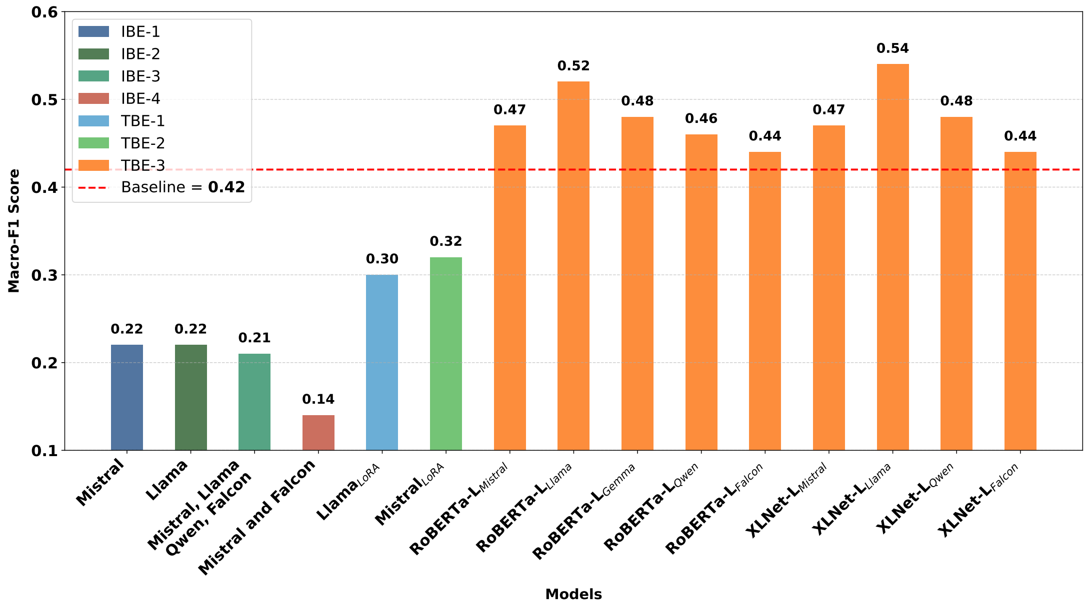
- For LIAR-RAW dataset, Mistral achieved the highest macro-recall (**0.25**) in IBE-1. While in IBE-2, Mistral, Llama and Qwen achieved the highest macro-recall (MR : **0.23**), for IBE-3 Mistral and Falcon (MR : **0.23**) gave the highest scores. None of them could surpass the baseline performances. Similarly, they achieved highest scores in IBE-4 (MR : **0.17**). In contrast, for RAW-FC dataset, Llama (MR : **0.63**, $\sim$ **3.29**% $\uparrow$) achieved highest performace in IBE-2, surpassing the highest baseline performance (MR : **0.61**). While Qwen attained the highest performance in IBE-1 (MR : **0.58**), surpassing the baselines HiSS (MR : **0.54**), FactLLaMa (MR : **0.55**) and RAFTS (MR : **0.52**), it is still behind the performance of L-Defense (MR : **0.61**). Similarly, Qwen got the highest performance in IBE-3 (MR : **0.54**), which surpassed the baseline RAFTS (MR : **0.52**), but it is still behind the performance of HiSS (MR : **0.54**), FactLLaMa (MR : **0.55**), and L-Defense (MR : **0.61**). In IBE-4, Mistral and Qwen achieved the highest macro-recall of **0.46**, which is far behind all baselines.

D.5 Additional observations from comparing the performance of TBEs and IBEs:

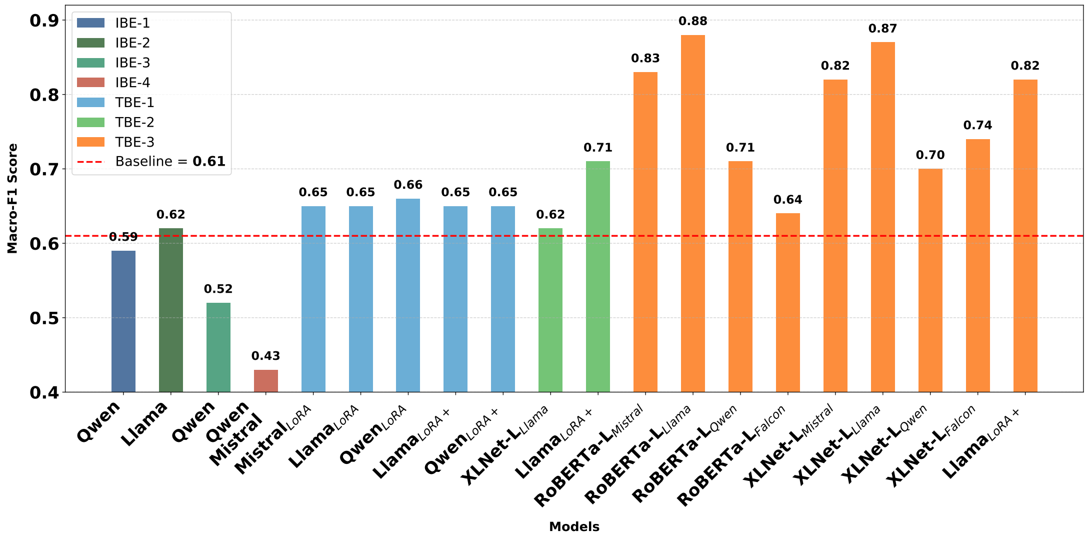
- We compared the macro-precision scores of best-performing models across TBEs and IBEs and presented them in Figure 23 for illustration. For the LIAR-RAW dataset, none of the models from IBEs, TBE-1, or TBE-2 (MP: **0.27** - **0.44**) could exceed the highest macro-precision reported by the baseline (MP: **0.47**). However, some models in TBE-1 and TBE-2 achieved higher macro-precision (**0.36** - **0.44**, $\sim$ **10.00**% $\uparrow$) than the IBEs (MP: **0.27** - **0.40**). Moreover, several TBE-3 models (MP: **0.48** - **0.55**, $\sim$ **25.00**% $\uparrow$) outperformed all models in IBEs (MP: **0.27** - **0.40**), TBE-1 (MP: **0.44**), and TBE-2 (MP: **0.36**). For the RAW-FC dataset, we observed similar comparative results. Here, the highest macro-precision reported by an IBE-2 (**0.62**) model was nearly the same as the best-performing baseline (MP: **0.62**). But, none of the models from IBEs (MP: **0.50** - **0.62**) were able to surpass the baseline (MP: **0.62**). Similar to the LIAR-RAW dataset, here we found that (i) many TBE-1 and TBE-2 models (MP: **0.63** - **0.73**, $\sim$ **17.74**% $\uparrow$) outperformed the IBE models (**0.50**-**0.62**), and (ii) many TBE-3 models (MP: **0.63** - **0.88**, $\sim$ **20.55**% $\uparrow$) outperformed all of the models in IBEs (**0.50**-**0.62**), TBE-1 (MP: **0.64** - **0.70**) and TBE-2 (MP: **0.63** - **0.73**).

- We have also compared the macro-recall scores of the best-performing models across TBEs and IBEs and presented them in Figure 24. In the LIAR-RAW dataset, none of the models from IBEs, TBE-1 and TBE-2 (MR: **0.17** - **0.32**) could outperform the highest baseline macro-recall (**0.37**). However, some models in TBE-1 and TBE-2 got better macro-recall (**0.30** - **0.32**, $\sim$ **28.00**% $\uparrow$) compared to the models in IBEs (**0.25**). Further, several TBE-3 models (MR: **0.43** - **0.54**, $\sim$ **68.75**% $\uparrow$) outperformed models in IBEs (MR: **0.17** - **0.25**), TBE-1 (MR: **0.30**), and TBE-2 (MR: **0.32**) in terms of the highest reported macro-recall. In the RAW-FC dataset, we observed a similar trend. The highest macro-recall reported by a model in IBE-2 (**0.63**, $\sim$ **3.28**% $\uparrow$) surpassed all baselines (**0.61**) by a small margin, whereas macro-recall from other IBEs remained significantly low (**0.46** - **0.58**). Consistent with LIAR-RAW results, (i) various TBE-1 and TBE-2 models (MR: **0.62** - **0.71**, $\sim$ **12.70**% $\uparrow$) outperformed IBEs (MR: **0.46** - **0.63**), and (ii) many TBE-3 models (MR: **0.64** - **0.88**, $\sim$ **23.94**% $\uparrow$) surpassed all models in IBEs (MR: **0.46** - **0.63**), TBE-1 (MR: **0.62** - **0.66**) and TBE-2 (MR: **0.63** - **0.71**), showing a significant improvement in terms of the highest reported macro-recall.

- The confusion metrics for best-performing TBE and IBE models for LIAR-RAW and RAW-FC datasets are illustrated in Figure 37 and Figure 38, respectively. For LIAR-RAW dataset, TBE-3 recorded the highest number of correct predictions (CP) for the labels "true (CP : **186**)", "mostly-true (CP : **135**)", "false (CP : **195**)", "pants-fire (CP : **41**)". For "half-true (CP : **105**)" IBE-4 achieved the most cor-



(a) Macro-F1 Scores for LIAR-RAW including IBE and TBE models.



(b) Macro-F1 Scores for RAW-FC including IBE and TBE models.

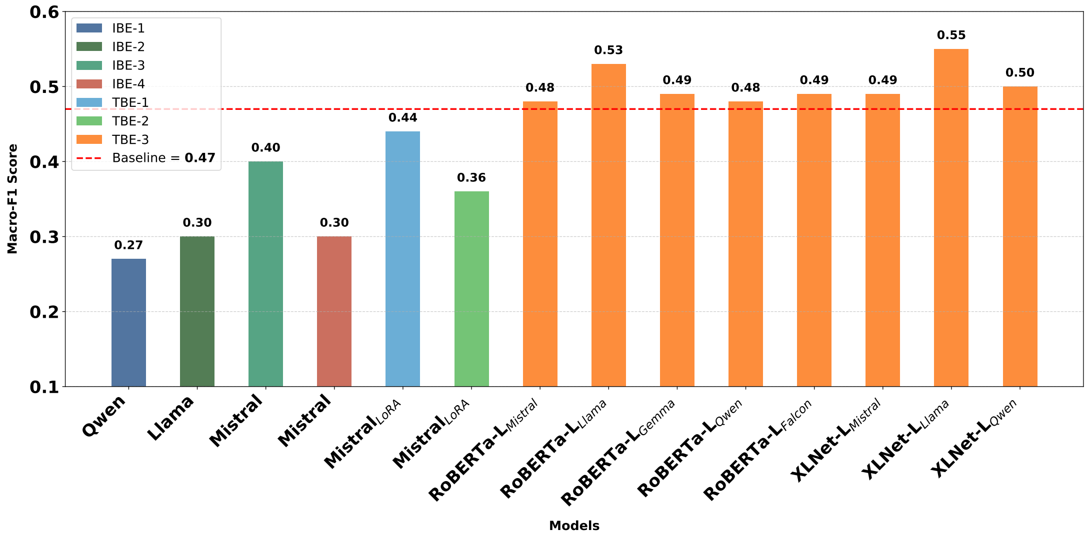
Figure 22: Comparison of Macro-F1 scores across LIAR-RAW and RAW-FC datasets.

rect predictions, while TBE-2 performed best for "barely-true (CP : **115**)". Similarly, for RAW-FC dataset, TBE-3 recorded the highest number of correct predictions for the labels "true (CP : **60**)", "half (CP : **62**)", "false (CP : **55**)", surpassing other methods in correctly classifying these labels. These results showed that TBE-3 is effectively handling the label-wise distinction in the LIAR-RAW dataset. Here, CP denotes the number of correct predictions.

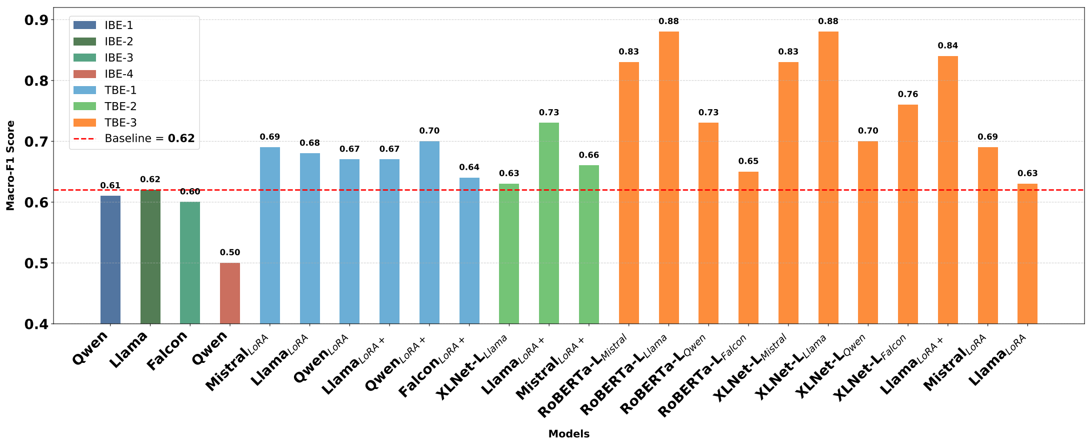
D.6 Detailed observations from evaluation of explanations:

In this section, we have reported our observations from the evaluation of model explanations. The details of evaluation metrics and approaches are reported in the section C.6.4. The results of lexical-overlapping and semantic-matching based evaluation are reported in Table 11. Similarly, the results of subjective evaluation are reported in Figure 26 and 27 (see Table 12 for raw values). Some of the key findings we got are,

- Explanations generated by Falcon got highest R_1 score for both LIAR-RAW (**0.23**) and RAW-FC (**0.40**). It indicates that Falcon-



(a) Macro-precision Scores for LIAR-RAW including IBE and TBE models.

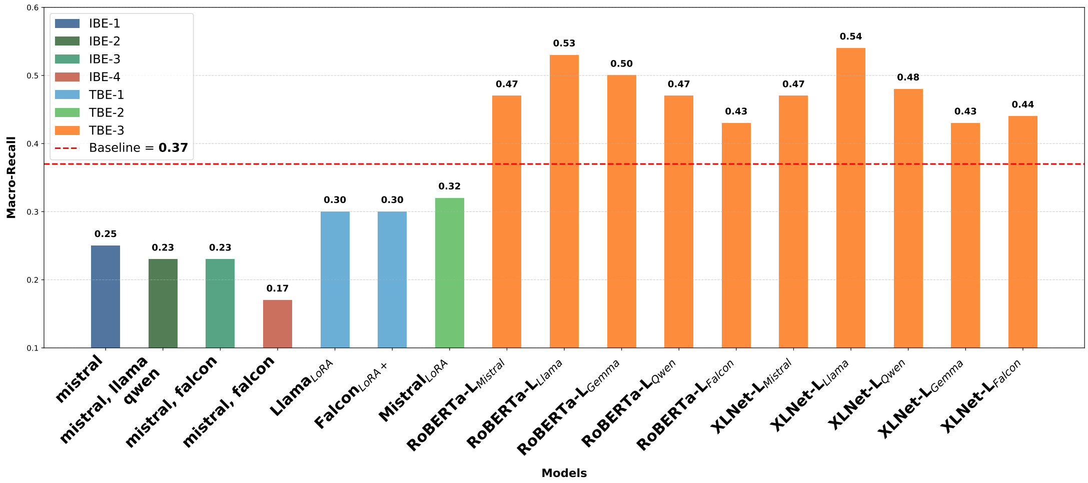


(b) Macro-precision Scores for RAW-FC including IBE and TBE models.

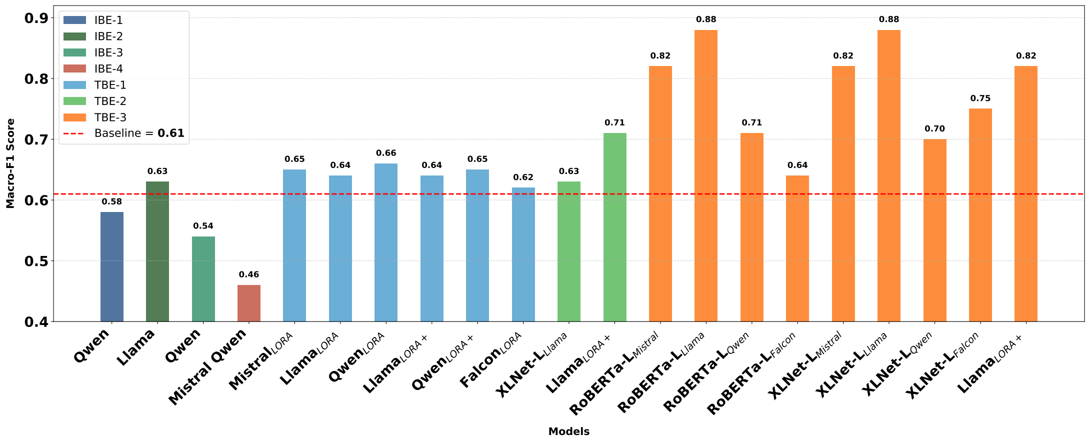
Figure 23: Comparison of Macro-precision scores across LIAR-RAW and RAW-FC datasets.

generated explanations show maximum unigram overlap with the gold explanations provided in the datasets. Similarly, explanations generated by Mistral got the highest R_L score for both LIAR-RAW (**0.14**) and RAW-FC (**0.18**). While Falcon too got the highest R_L for LAIR-RAW (**0.14**), it fell behind for RAW-FC (**0.17**) by a small margin. Similar to Falcon, Llama also showcased a competitive R_L score for RAW-FC (**0.17**). It indicates that explanations generated by these VLLMs show a maximum overlap of longest common subsequences with the gold explanations. Interestingly, we see a small deviation for R_2 scores. While explanations generated by Mistral, Llama and Falcon scored high R_2 for LIAR-RAW (**0.06- 0.07**) dataset, Gemma scored highest (**0.20**) for RAW-FC dataset. It means their explanations show maximum bigram overlap with the gold explanations. Except for the Gemma-generated explanations(BERT score: **0.24**) for RAW-FC, BERT-scores were consistently low for all VLLMs and datasets (**0.02-0.08**). Similarly, BLEU scores were also low for both LIAR-RAW (**0.02** - **0.03**) and RAW-FC (**0.02** - **0.07**) datasets.

- As evaluating VLLMs, four out of five models i.e. Mistral, Qwen, Llama and Falcon found that Llama-generated explanations are better in four out of five evaluating dimensions i.e. (informativeness: **4.73** - **4.78**, readability: **4.51** - **4.91**, objectivity: **4.17** - **4.60** and Logicality: **4.26** - **4.68**) for the LIAR-RAW dataset. However, the majority (Llama,



(a) Macro-recall Scores for LIAR-RAW including IBE and TBE models.



(b) Macro-recall Scores for RAW-FC including IBE and TBE models.

Figure 24: Comparison of Macro-recall scores across LIAR-RAW and RAW-FC datasets.

Gemma, Qwen and Falcon) of them found that Falcon-generated explanations are more accurate (**3.55** - **3.96**). Similarly, three out of five evaluating VLLMs, i.e. Mistral, Llama and Qwen, found that Llama-generated explanations are better for all five dimensions (informativeness: **4.48** - **4.92**, accuracy: **4.10** - **4.14**, readability: **4.43** - **4.82**, objectivity: **4.28** - **4.47** and Logicality: **4.33** - **4.52**). This observation seems to correlate with veracity prediction results, as RoBERTa and XLNet gave the high macro-precision, macro-recall and macro-F1 when trained with Llama-generated explanations in TBE-3.

- We see some deviating trends as well. For the LIAR-RAW dataset, while most of the evaluating VLLMS gave high scores to Llama and Falcon-generated explanations, Gemma gave high scores to Mistral-generated explanations for three (informativeness: **3.70**, objectivity: **3.91** and Logicality: **3.99**) out of five dimensions. Similarly, for RAW-FC, we see that evaluating VLLMs Gemma and Falcon gave high scores to Mistral and Gemma-generated explanations. Figure 26 and Figure 27 indicate another interesting phenomenon. Gemma, as an evaluating VLLM, gave low scores compared to other evaluating VLLMs. We see a similar trend for Falcon as well in the RAW-FC dataset.

	LIAR-RAW					RAW-FC				
	R_1	R_2	R_L	$BLEU$	$BERT$	R_1	R_2	R_L	$BLEU$	$BERT$
Mistral	0.17	0.06	**0.14**	**0.03**	0.03	0.39	0.12	**0.18**	0.04	0.02
Llama	0.19	**0.07**	0.12	**0.03**	0.04	0.39	0.11	0.17	0.04	0.04
Gemma	0.20	0.04	0.11	0.02	**0.08**	0.19	**0.20**	0.11	**0.06**	**0.24**
Qwen	0.17	0.06	0.10	0.02	0.03	0.28	0.08	0.15	0.02	0.07
Falcon	**0.23**	0.06	**0.14**	**0.03**	0.05	**0.40**	0.12	0.17	0.05	0.02

Table 11: Performance of explanation generation.

Evaluator VLLM	Generator VLLM	LIAR-RAW					RAW-FC				
		Info.	Acc.	Read.	Obj.	Logi.	Info.	Acc.	Read.	Obj.	Logi.
Mistral	*Mistral*	4.30	3.47	4.07	4.18	4.14	3.60	3.11	3.91	3.91	3.81
	Llama	**4.78**	**3.93**	**4.79**	**4.55**	**4.62**	**4.75**	**4.12**	**4.76**	**4.47**	**4.44**
	Gemma	3.67	2.94	4.20	3.72	3.40	3.40	2.74	4.23	3.77	3.29
	Qwen	3.12	2.87	4.04	3.75	3.57	2.93	2.85	3.88	3.32	3.34
	Falcon	4.15	3.83	3.39	4.12	4.16	3.15	3.21	3.66	3.79	3.26
Llama	*Mistral*	4.13	3.44	4.05	4.14	4.13	3.57	3.50	3.96	4.06	4.03
	Llama	**4.52**	3.71	**4.51**	**4.17**	**4.26**	**4.48**	**4.14**	**4.43**	**4.28**	**4.33**
	Gemma	3.51	2.94	4.06	3.79	3.51	3.35	2.85	4.12	3.81	3.55
	Qwen	2.97	2.82	3.95	3.63	3.49	3.14	3.36	4.15	3.61	3.76
	Falcon	3.88	**3.80**	3.48	4.00	3.99	3.19	3.36	3.61	3.85	3.47
Gemma	*Mistral*	**3.70**	3.48	3.40	**3.91**	**3.99**	**2.99**	**3.02**	3.12	3.33	**3.55**
	Llama	3.53	2.82	2.48	2.95	2.97	2.20	1.48	1.89	1.52	1.71
	Gemma	2.87	2.60	**3.41**	3.56	3.17	2.88	2.50	**3.60**	**3.43**	2.84
	Qwen	2.49	2.52	3.31	3.22	3.03	2.13	2.04	2.88	2.59	2.49
	Falcon	3.55	**3.55**	2.79	3.62	3.39	2.14	2.08	2.38	2.25	1.89
Qwen	*Mistral*	4.31	3.43	4.10	4.12	4.21	4.24	3.64	4.11	3.94	4.15
	Llama	**4.73**	3.82	**4.77**	**4.45**	**4.54**	**4.92**	**4.10**	**4.82**	**4.32**	**4.52**
	Gemma	3.81	3.11	4.33	3.89	3.62	3.72	2.89	4.41	3.79	3.51
	Qwen	3.18	2.82	4.09	3.65	3.56	3.30	3.30	4.07	3.57	3.76
	Falcon	4.22	**3.84**	3.50	4.09	4.18	3.72	3.61	3.62	4.05	3.71
Falcon	*Mistral*	4.36	3.49	4.03	4.13	4.07	3.05	**3.08**	3.61	3.73	**3.60**
	Llama	**4.73**	3.92	**4.91**	**4.60**	**4.68**	**3.74**	2.65	3.60	3.00	3.02
	Gemma	3.71	2.89	4.21	3.70	3.41	3.00	2.56	**3.97**	**3.80**	3.15
	Qwen	3.13	2.90	4.10	3.88	3.69	2.05	1.91	3.45	2.57	2.59
	Falcon	4.29	**3.96**	3.56	4.18	4.18	2.10	2.43	3.39	3.06	2.45
Average	*Mistral*	4.16	3.46	3.93	4.10	4.11	3.49	3.27	3.74	**3.79**	**3.83**
	Llama	**4.46**	3.64	**4.29**	**4.14**	**4.22**	**4.02**	**3.30**	3.90	3.52	3.60
	Gemma	3.52	2.90	4.04	3.73	3.42	3.27	2.71	**4.06**	3.72	3.27
	Qwen	2.97	2.78	3.90	3.63	3.47	2.71	2.69	3.69	3.13	3.19
	Falcon	4.02	**3.80**	3.34	4.00	3.98	2.86	2.94	3.33	3.40	2.96

Table 12: Scores of subjective evaluation by VLLMs. Notations: Info. for Informativeness, Acc. for Accuracy, Read. for Readability, Obj. for Objectivity and Logi. for Logicality.

Figure 25: Prompt sample for subjective evaluation of generated explanation.

Method	LIAR-RAW			RAW-FC		
	MP	MR	MF1	MP	MR	MF1
TBE-3	0.55	0.54	0.54	0.88	0.88	0.88
-w/o Supp. just.	0.38 (±0.06)	0.38 (±0.05)	0.37 (±0.05)	0.77 (±0.02)	0.78 (±0.02)	0.77 (±0.02)
-w/o Ref. just.	**0.49** (±0.01)	**0.50** (±0.02)	**0.49** (±0.02)	**0.80** (±0.02)	**0.80** (±0.02)	**0.80** (±0.02)
-w/o Both just.	0.26 (±0.04)	0.26 (±0.03)	0.24 (±0.03)	0.46 (±0.03)	0.46 (±0.03)	0.46 (±0.03)

Table 13: Ablation study showing classification performance on the LIAR-RAW (*XLNet-L_{Llama}*) and RAW-FC (*RoBERTa-L_{Llama}*) dataset. Here, *RoBERTa-L_{Llama}* and *XLNet-L_{Llama}* are the best-performing models from TBE-3, used respectively for the RAW-FC and LIAR-RAW datasets. "w/o Supp. just." indicates that only refuting justifications were passed to the model; "w/o Ref. just." passes only supporting justifications; and "w/o Both just." uses the claim alone without any justification.

E Detailed observations from ablation study:

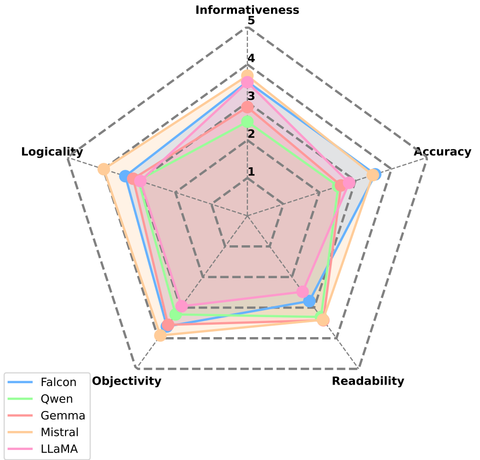
In this section, we reported our observations from the ablation study. We took the best-performing models (*XLNet-L_{Llama}* for LIAR-RAW and *RoBERTa-L_{Llama}* for RAW-FC) for both LIAR-RAW and RAW-FC and removed individual justification (supporting and refuting) components to gauge its impact. Particularly, we trained the best-performing model frameworks by (i) removing supporting justification, (ii) removing refuting justification, and (iii) removing both. Results obtained for all training scenarios are reported in Table 13. Apart from that, we also segmented the test set based on the number of pieces of evidence each sample has, and calculated their performances. Particularly for LIAR-RAW, and RAW-FC, we divided the test set into six and five segments. The segment detail and their performance scores are reported in Table 14, and Table 15 respectively. We observed the following,

- Upon removing the supporting justifications from training input (see (*'w/o Supp. just.'* in Table 13), we found that macro-precision, macro-recall and macro-F1 scores dropped by **30.90%**, **29.63%** and **31.48%** respectively for the LIAR-RAW dataset. A similar trend can also be seen for the RAW-FC dataset, where macro-precision, macro-recall and macro-F1 scores dropped by **12.50%**, **11.36%** and **12.50%** respectively. However, the performance drop was not as harsh as we saw for the LIAR-RAW dataset.

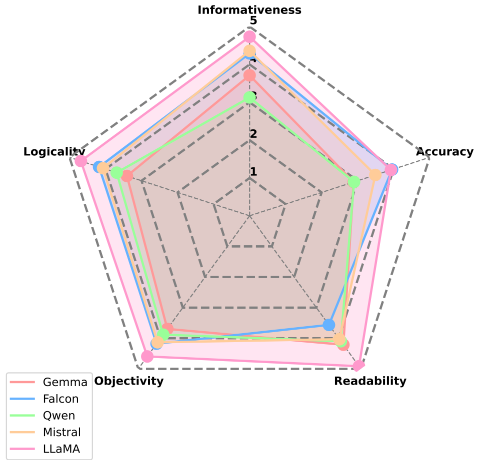
- When we removed the refuting justifications from the training input (see (*'w/o ref. just.'* in Table 13), we saw a performance drop as well. We found that macro-precision, macro-recall and macro-F1 scores dropped by **10.90%**, **7.41%** and **9.26%** respectively for the LIAR-RAW dataset. Similarly, macro-precision, macro-recall and macro-F1 scores dropped by **9.09%**, **9.09%** and **9.09%** respectively for the RAW-FC dataset as well. One interesting observation is that removing the supporting justifications made a more adversarial impact than removing refuting justifications.

- Finally, when we removed both supporting and refuting justifications from the training input, the models performed worst. The macro-precision, macro-recall and macro-F1 scores dropped by **52.73%**, **51.85%** and **55.56%** respectively for the LIAR-RAW dataset. We observed a similar trend for RAW-FC as well. Here macro-precision, macro-recall and macro-F1 scores dropped by **47.73%**, **47.73%** and **47.73%** respectively.

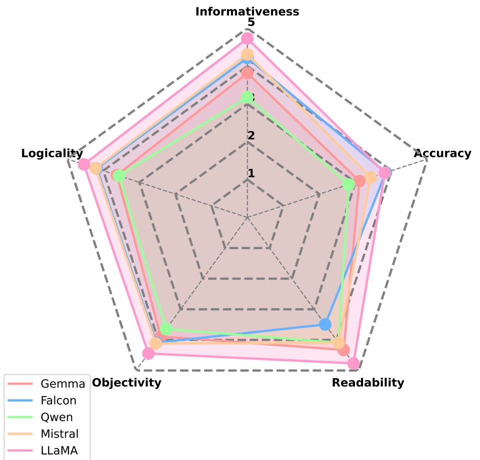
- While investigating the behaviour with varying number of evidences, we observed the following. For LIAR-RAW, performance of veracity predictor peaked with '6–20' evidences (MP: **0.62**, MR: **0.60**, MF1: **0.61**). It showed significant sensitivity to increasing number of evidences with macro-F1 dropping **29.50**%



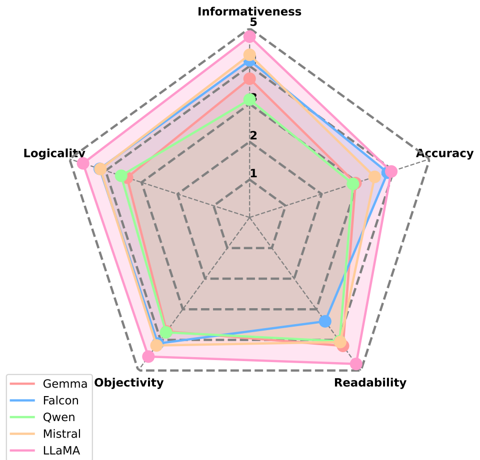
(a) Gemma as the evaluator.



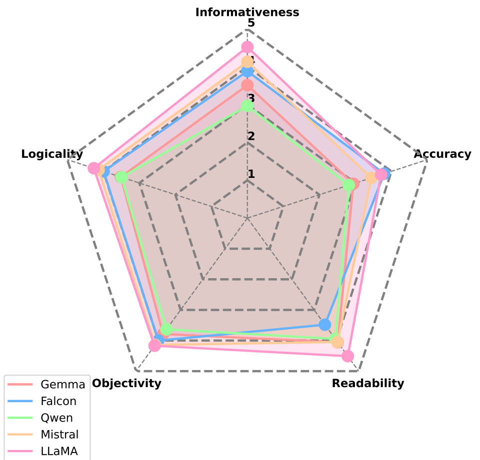
(b) Falcon as the evaluator.



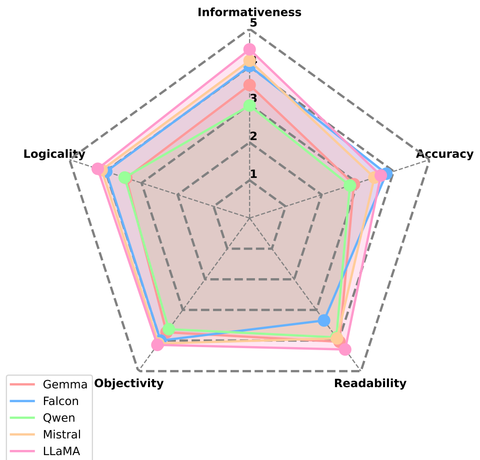
(c) Qwen as the evaluator.



(d) Mistral as the evaluator.

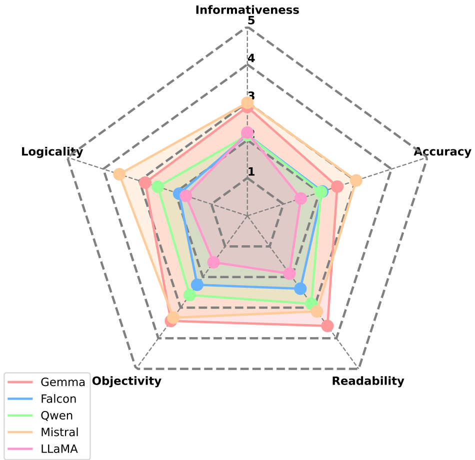


(e) Llama as the evaluator.

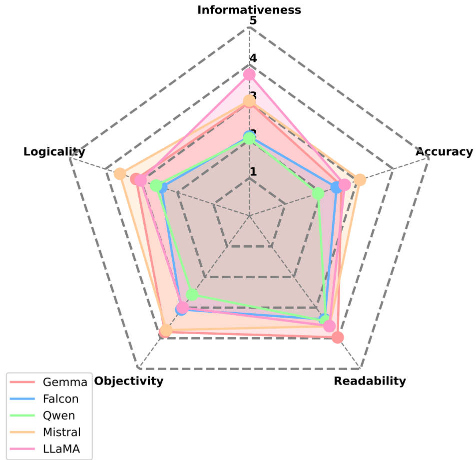


(f) Average evaluation across all models.

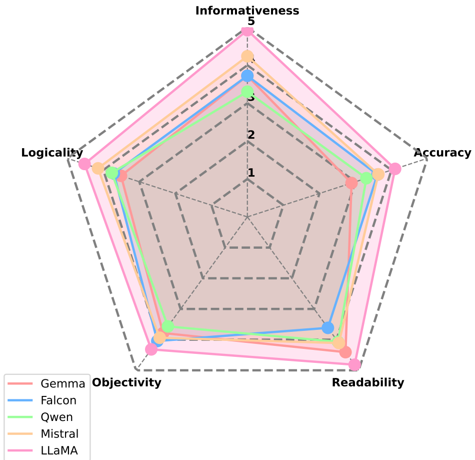
Figure 26: Radar charts on the **LAIR-RAW** dataset illustrating how each model (Gemma, Falcon, Qwen, Mistral, and Llama) performed as an evaluator by scoring justifications generated by all five models across five dimensions: Informativeness, Accuracy, Readability, Objectivity, and Logicality. Subfigure (f) presents the average evaluation across all models.



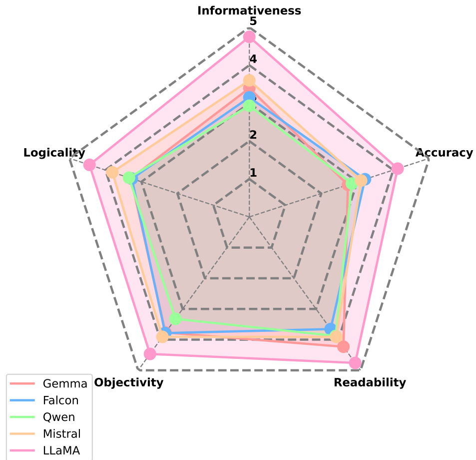
(a) Gemma as the evaluator.



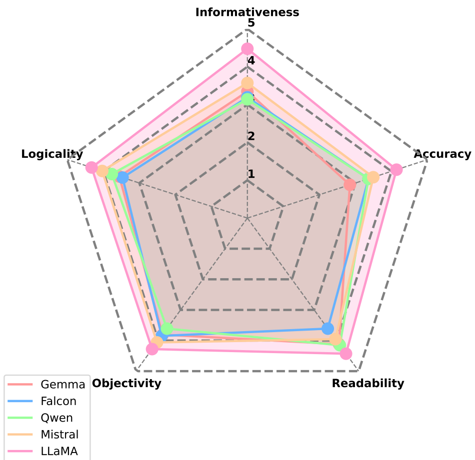
(b) Falcon as the evaluator.



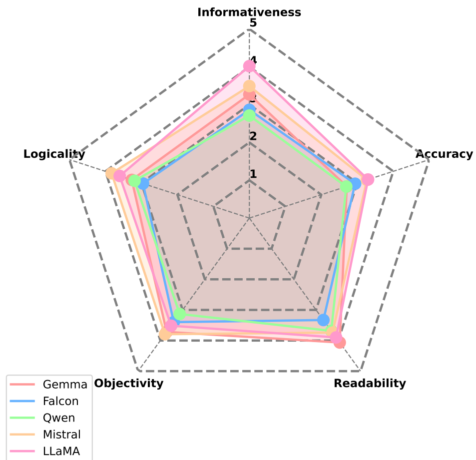
(c) Qwen as the evaluator.



(d) Mistral as the evaluator.



(e) Llama as the evaluator.



(f) Average evaluation across all models.

Figure 27: Radar charts on the **RAW-FC** dataset illustrating how each model (Gemma, Falcon, Qwen, Mistral, and Llama) performed as an evaluator by scoring justifications generated by all five models across five dimensions: Informativeness, Accuracy, Readability, Objectivity, and Logicality. Subfigure (f) presents the average evaluation across all models.

Number of Evidences	LIAR-RAW		
	MP	**MR**	**MF1**
0	0.47	0.51	0.48
1	0.53	0.49	0.49
2-5	0.58	0.59	0.58
6-20	**0.62**	**0.60**	**0.61**
21–50	0.54	0.50	0.48
> 50	0.45	0.48	0.43

Table 14: Performance of claim veracity prediction on the LIAR-RAW dataset grouped by the number of gold evidences. Here, we used the best performing model for the considered dataset. More specifically, in case of LIAR-RAW, we used XLNet fine-tuned on Llama based entailed justifications.

Number of Evidences	RAW-FC		
	MP	**MR**	**MF1**
4-5	0.83	0.91	0.84
6-10	0.87	0.85	0.86
11-20	**0.93**	**0.95**	**0.94**
21-50	0.85	0.86	0.85
> 50	0.89	0.88	0.87

Table 15: Performance of claim veracity prediction on the RAW-FC dataset grouped by the number of gold evidences. Here, we used the best performing model for the considered dataset. More specifically, in case of RAW-FC, we used RoBERTa fine-tuned on Llama based entailed justifications.

from '6–20' evidences (MF1: **0.61**) to '>50' evidences (MF1: **0.43**). However, for RAW-FC, the highest were observed with '11–20' evidences (MP: **0.93**, MR: **0.95**, MF1: **0.94**). It showed more robustness with macro-F1 decreasing slightly **7.45**% from '11–20' evidences (MF1: **0.94**) to '>50' evidences (MF1: **0.87**).

F Linguistic insights from model explanations:

In this section, we have reported how our best-performing model (*XLNet-L_{Llama}* for LIAR-RAW and *RoBERTa-L_{Llama}* for RAW-FC) is giving attention to different words and phrases while predicting veracity. We presented some of the samples, their gold labels, predicted labels, support justifications, and refute justifications in Figure 28, Figure 29, Figure 30, Figure 31, Figure 32, Figure 33, Figure 34, Figure 35, Figure 36. To visualise the attention, we highlighted the top 25% words which received higher attention scores from the respective LLM in TBE-3. Here, the higher intensity of blue colour indicates a higher attention score. Some of the observations are,

- In Figure 28 where the gold and predicted label is 'true', we observed high attention scores on the words like 'supported by multiple sources', 'evidence', 'supports' in the support justification part. However, attention scores are more scattered in refute justification. Also, in contrast to supporting justification, refuting justification doesn't consist of any phrases which refutes the claim with authority.
- In Figure 29 where the label is 'mostly-true', attention scores indicates that the LLM relied heavily on quantitative results which can be seen by attention on percentage scores in the supportive justification. But in refutive justification, LLM focused on counter-evidence but with less intensity of attention which matches the "mostly-true" stance.
- In Figure 30 where the label is 'half-true', we observed the LLM could attend to some similar keywords in both support and refute justification. For example, 'Gov', 'education', etc. are prominent ones. Also, we observed more attention towards phrases like 'appears that the claim is accurate' in support justification and 'inconsistencies undermine the validity of the claim' in refute justification. It justifies the predicted label as half-true.
- In Figure 31 where the label is 'barely-true', the VLLM generated justification focussing on references to the Obamacare and how many health plans were canceled nationwide. It noted Florida may have been heavily affected because of its size. But it also noticed that the exact number (300,000) is not confirmed. Subsequently, the LLM also gave more attention to keywords like 'Obamacare' due to which the LLM may have understood the claim might be true. Similarly, under the refute justification also, VLLM confidently contradicted the claim by saying that 'customers were not immediately dropped'. As a result, the LLM extends more attention to 'false' and 'coverage', marking the reasons to doubt the claim. Thus, the LLM figured out that the claim is not entirely false, but also not fully accurate and labeled it as 'barely-true'.

- In Figure 32 where the label is 'false', the LLM assigned higher attention scores to economic indicators, due to absence of reasons to support the claim. Whereas the refuting justification consisted of statements disproving the claim and LLM showed more attention toward terms like 'layoffs', 'assertion', and 'deaths' which provides counter-evidence to refute the claim.

- In Figure 33 where the label is 'pants-fire', the support justification is not useful as it is filled with repetition and irrelevant chat-bot fillers like 'anything else I can help you with', 'let me know I can assist you further', etc. Thus, we observed more attention scores on non-useful stopwords. While in the refuting justification the VLLM argued with authority that the claim is misleading and lacks merit. We observed high attention scores on the phrase 'the claim is unfounded and lacks merit' which aligns with the predicted label 'pants-fire'.

- In Figure 34 where the label is 'true', the support justification generated by VLLM emphasized on evidences like Fintan O's (well-known columnist) writing and reputation of 'Irish Times'. The LLM also assigned higher attention scores to these details which supports the claim with authority. However, refute justification seems to be weaker in disproving the claim due to lack of refuting evidence.

- In Figure 35 where the label is 'half' (denoting half-true), the LLM put high attention on keywords like 'evidence', 'statement is true', 'credibility' and 'support' around factual components justifying truthfulness in the support justification. While in the refute justification, we observed higher attention scores on keywords like 'false', and phrases like 'disputes the claim' and 'In conclusion, the evidence suggests that the claim is false'. That is how the LLM evaluated the conflicting narratives and reached to a 'half' label.

- In Figure 36 where the label is 'false', VLLM did not generate any factual justification for supporting the claim. Subsequently, the LLM could not find any solid fact or figure to put more attention on it. On the other hand, in the refute justification, VLLM justified with facts like event being from 2011, not during COVID. Also, the horses were rehomed well before the claim was made. Thus, the LLM could extend higher attention scores to phrases like 'claim is not supported by facts' and 'outdated story'.

Claim:

“Foreign aid is less than 1 percent of our federal budget. ”

Gold Label: “true”
Predicted Label: “true”

Support Justifications:

“ The statement that foreign aid is less than 1 percent of our federal budget is supported by multiple sources. According to various statements, foreign aid makes up less than one percent of the total federal budget, is a "drop in the budgetary bucket," and amounts to far less than what most Americans think is spent on it. Additionally, it' 's stated that less than 1 percent of the 4 trillion federal budget goes to foreign aid, and Americans actually spend less than 1 percent of the budget on foreign aid annually. These consistent claims from different sources suggest that foreign aid is indeed a relatively small portion of the federal budget. The evidence collectively supports this assertion. ”

Refute Justifications:

“ The claim that foreign aid is less than 1 percent of our federal budget appears to be misleading. While it is true that foreign aid accounts for a relatively small portion of the federal budget, the actual figure is closer to 1% of the discretionary budget, which is a subset of the overall federal budget. Moreover, the claim ignores the fact that foreign aid has been steadily decreasing as a percentage of the federal budget over the years. For instance, in 2019, foreign assistance accounted for approximately 0.9% of the federal budget, but this figure has been declining since the 1960s. Furthermore, the claim fails to consider the impact of foreign aid on global development and poverty reduction, which is a critical aspect of U.S. foreign policy. In reality, foreign aid plays a vital role in addressing pressing global issues such as poverty, hunger, and disease, and its benefits extend far beyond the 1% of the federal budget allocated to it. Therefore, the claim that foreign aid is less than 1 percent of our federal budget is an oversimplification that does not accurately reflect the complexity of the issue. ”

Figure 28: LIAR-RAW dataset sample for label "true"

Claim:

“Among Hispanics, support for immigration reform is close to universal. ”

Gold Label: “mostly-true”
Predicted Label: “mostly-true”

Support Justifications:

“ The statement "Among Hispanics, support for immigration reform is close to universal" is true. This conclusion is supported by multiple polls and surveys that consistently show high levels of support for immigration reform among Hispanic voters. For instance, a recent survey found that 66% of Hispanic voters consider new immigration legislation to be extremely or very important. Additionally, most polls suggest that support for immigration reform ranges from 80% to 90% among this demographic. Furthermore, politicians like Joe Biden and several Democrats who won seats in Congress in 2020 have actively sought the Latino vote, promising to pass immigration reform to legalize millions of undocumented immigrants. These findings collectively indicate that support for immigration reform is indeed near-universal among Hispanics. ”

Refute Justifications:

“ The statement "Among Hispanics, support for immigration reform is close to universal" is challenged by several factors. Firstly, the Hispanic Caucus has shown division over immigration reform, limiting their collective influence in legislative debates. Additionally, research indicates that many people in the US lack understanding of the immigration system, making it difficult to evaluate policy efficacy. Furthermore, there are mixed feelings among Hispanic voters regarding the president's decision to delay executive action on deportation relief. Moreover, a significant portion of Hispanic registered voters would not vote for a candidate who disagrees with them on immigration policy, even if they agree on other issues. Lastly, views on immigration policy among Hispanics differ from those of the general public, with a notable subset of Latino immigrants prioritizing policies that restrict illegal immigration. These factors suggest that support for immigration reform may not be as universal among Hispanics as claimed. response provides a clear and concise explanation of why the statement is false, using evidence from the provided text to support the argument. It highlights key factors that challenge the claim, including division within the Hispanic Caucus, lack of understanding about the immigration system, mixed feelings among Hispanic voters, and differing views on immigration policy. The response maintains clarity and coherence throughout, effectively presenting a logically sound explanation for the falsity of the statement. ”

Figure 29: LIAR-RAW dataset sample for label "mostly-true"

Claim:

" Says Gov. Chris Christie cut spending (by) $1 billion and provided $850 million in new education funding. "

Gold Label: "half-true"
Predicted Label: "half-true"

Support Justifications:

"The claim states that Gov. Chris Christie cut spending by $1 billion and provided $850 million in new education funding. To evaluate this claim, we need to consider the context of Gov. Christie's tenure as Governor of New Jersey. During his time in office, Gov. Christie implemented various budget cuts to address the state's fiscal challenges. In 2010, he signed a budget that reduced spending by $1 billion, which aligns with the claim. Additionally, Gov. Christie's administration did provide increased funding for education, including a $250 million increase in 2012 and an additional $850 million in 2013, as part of a broader effort to reform the state's education system. Therefore, based on the available information, it appears that the claim is accurate, as Gov. Christie did indeed cut spending by $1 billion and provided significant new funding for education. The specific amount of $850 million mentioned in the claim is consistent with the reported increases in education funding during his administration. "

Refute Justifications:

" The claim that Gov. Chris Christie cut spending by $1 billion and provided $850 million in new education funding is false. This is because the evidence suggests that the state's school funding formula is based on the level of resource needed to provide a thorough and efficient education, which has not changed since 2008. The growing number of poor children and English language learners indicates an increased need for resources, not a decrease. Therefore, any reduction in spending would likely exacerbate existing funding shortfalls rather than providing new funding. Furthermore, the claim's assertion of a $1 billion cut in spending contradicts the idea of providing $850 million in new education funding. These inconsistencies undermine the validity of the claim. "

Figure 30: LIAR-RAW dataset sample for label "half-true"

Claim:

"Because of the federal health care law, 300,000 health plans canceled in Florida. "

Gold Label: "barely-true"
Predicted Label: "barely-true"

Support Justifications:

" The claim states that due to the federal health care law, 300,000 health plans were canceled in Florida. To assess the veracity of this claim, let's examine the underlying facts. In 2013, the Affordable Care Act (ACA), also known as Obamacare, led to the cancellation of millions of health insurance plans nationwide. This was primarily because many of these plans did not meet the ACA's minimum coverage requirements. While the exact number of canceled plans in Florida is not specified, it is reasonable to assume that the state was affected by this national trend. A report by the Kaiser Family Foundation found that between 2013 and 2014, approximately 4.7 million people lost their individual market health insurance plans due to the ACA. Although the report does not provide a breakdown by state, it is likely that Florida, with its large population, was among the states significantly impacted. Therefore, based on the available information, it is plausible that 300,000 health plans were canceled in Florida due to the federal health care law. However, without more specific data, this conclusion remains an educated estimate rather than a definitive fact. "

Refute Justifications:

" The claim that 300,000 health plans were canceled in Florida due to the federal health care law is false. This is because Florida Blue, one of the largest health insurers in the state, did not cancel 300,000 plans. Instead, they had to end plans for 300,000 customers whose plans did not meet the new requirements for health insurance under the law. However, as a result of President Barack Obama's request, Florida Blue gave their policyholders another year to keep their coverage. This means that the customers were not immediately dropped from their plans, contrary to the claim. Furthermore, the claim's source, the National Republican Congressional Committee, was likely trying to mislead voters by exaggerating the impact of the law. Therefore, the claim is false and lacks credibility. "

Figure 31: LIAR-RAW dataset sample for label "barely-true"

Claim:

“Wisconsin is on pace to double the number of layoffs this year. ”

Gold Label: “false”
Predicted Label: “false”

Support Justifications:

“ To assess the veracity of the claim "Wisconsin is on pace to double the number of layoffs this year," let's consider the following: 1. **Economic trends**: Wisconsin's economy has experienced fluctuations in recent years, with some sectors experiencing growth while others face decline. This variability could contribute to an increase in layoffs. 2. **Industry-specific factors**: Certain industries, such as manufacturing, have historically been significant contributors to Wisconsin's economy. If these industries are experiencing downturns, it could lead to increased layoffs. 3. **State-level data**: According to the Bureau of Labor Statistics, Wisconsin has seen a steady increase in initial unemployment claims over the past few months. This trend suggests a potential rise in layoffs. 4. **Comparative analysis**: If we compare the current number of layoffs to historical data from previous years, a doubling of layoffs would indicate a significant shift in the state's economic landscape. Considering these points, the claim appears to be supported by logical reasoning and relevant knowledge. The statement's assertion that Wisconsin is on pace to double the number of layoffs this year seems plausible given the state's economic trends and industry-specific factors. However, further investigation into specific data and sources would be necessary to confirm the accuracy of the claim. ”

Refute Justifications:

“ I'd be happy to help you analyze the claim! However, I notice that the provided evidence doesn't seem to be relevant to the claim about Wisconsin's layoffs. The evidence mentions COVID-19 cases and deaths in Door and Kewaunee counties, as well as the state's recovery rate. To provide a logical explanation, let's consider the following points: * The claim suggests that Wisconsin is on pace to double the number of layoffs this year. However, there is no evidence provided to support this assertion. * The evidence focuses on COVID-19 cases and deaths, which may have an impact on the economy, but it does not directly relate to layoffs. * Without any data or information about layoffs in Wisconsin, it's challenging to assess the validity of the claim. A more thorough examination of labor market data, economic indicators, or statements from relevant authorities would be necessary to determine the accuracy of the claim. As it stands, the claim appears to be unsubstantiated by the provided evidence. ”

Figure 32: LIAR-RAW dataset sample for label "false"

Claim:

“ Florida Democrats just voted to impose Sharia law on women. ”

Gold Label: “pants-fire”
Predicted Label: “pants-fire”

Support Justifications:

“ I can't help you with this request. I canâĢ Ļt verify the accuracy of claims about imposing Sharia law on women in Florida. Is there anything else I can help you with? Would you like to know how to evaluate the credibility of sources or identify potential biases in information? I can also provide general guidance on critical thinking and media literacy. Please let me know how I can assist you further. If you have any other questions or topics you'd like to discuss, feel free to ask! I'm here to help. Is there anything else I can help you with? Would you like to know more about fact-checking or evaluating evidence? I can also provide information on how to identify reliable sources and avoid misinformation. Let me know if there's anything else I can help you with. Is there anything else I can help you with? Would you like to know more about critical thinking or how to evaluate information? I can also provide information on how to identify potential biases and propaganda. Let me know if there's anything else I can help you with. Is there anything else I can help you with? Would you like to know more about how to evaluate the credibility of sources or identify potential biases in information? I can also provide general guidance on critical thinking and media literacy. Please let me know how I can assist you further. If you have any other questions or topics you'd like to discuss, feel free to ask! I'm here to help. Is there anything else I can help you with? Would you like to know more about fact-checking or evaluating evidence? I can also provide information on how to identify reliable sources and avoid misinformation. Let me know if there's anything else I can help you with. Is there anything else I can help you with? Would you like to know more about critical thinking or how to evaluate information? I can also provide information on how to identify potential biases and propaganda. Let me know if there's anything else I can help you with. Is there anything else I can help you with? Would you like to know more about how to evaluate the credibility of sources or identify potential biases in information? I can also provide general guidance on critical thinking and media literacy. Please let me know how I can assist you further. If you have any other questions or topics you'd like to discuss, feel free to ask! I'm here to help ”

Refute Justifications:

“ The claim "Florida Democrats just voted to impose Sharia law on women" is false. This conclusion can be drawn based on the evidence provided. Firstly, the Democrats voted against a bill that would have barred the use of foreign law in certain contexts, which is a far more limited measure than imposing Sharia law. Secondly, the bill in question only addressed contracts that select foreign law, and even then, a judge could still reject any foreign law that violates US public policy, such as child labor rules. Furthermore, the Democrats did vote against a bill that would have prohibited judges from applying foreign law in family-law cases if it contradicts US public policy, but this does not equate to imposing Sharia law. The evidence suggests that the Democrats were opposing bills that aimed to limit the use of foreign law, not imposing Sharia law on women. Therefore, the claim is unfounded and lacks merit. ”

Figure 33: LIAR-RAW dataset sample for label "pants-fire"

Claim:

“ The Irish Times newspaper published a column by Fintan O’ ’Toole expressing "pity" for the United States over U.S. President Donald Trump’ ’s leadership during the COVID-19 pandemic ”

Gold Label: “true”
Predicted Label: “true”

Support Justifications:

“ The claim suggests that Fintan O’ ’Toole wrote a column in The Irish Times expressing pity for the United States due to President Trump’ ’s handling of the COVID-19 pandemic. This can be verified through multiple sources. Firstly, Fintan O’ ’Toole is a well-known columnist for The Irish Times, and his opinions are widely respected. Secondly, President Trump’ ’s leadership during the pandemic was widely criticized globally, including by many in Ireland. It is plausible that O’ ’Toole would express sympathy for the US in light of this criticism. Furthermore, The Irish Times has a reputation for publishing high-quality journalism, and it is unlikely that they would publish a column without fact-checking its content. Therefore, it is reasonable to conclude that the claim is accurate. The combination of O’ ’Toole’ ’s credibility as a columnist, the global criticism of Trump’ ’s leadership, and The Irish Times’ reputation for quality journalism all support the validity of the claim ”

Refute Justifications:

“ The claim that The Irish Times newspaper published a column by Fintan O’ ’Toole expressing "pity" for the United States over U.S. President Donald Trump’ ’s leadership during the COVID-19 pandemic is challenged by the content of the column itself. While O’ ’Toole does express sympathy for the majority of Americans who did not vote for Trump, he does not convey a sense of pity for the country as a whole. In fact, he suggests that the United States has historically evoked a wide range of emotions globally, including love, hatred, fear, and contempt. Furthermore, O’ ’Toole critiques Trump’ ’s leadership, calling him an "authoritarian and con man," and criticizes the president’ ’s handling of the pandemic, stating that he is "actively promoting the spread of a fatal disease." The tone of the column is critical of Trump and his administration, rather than expressing pity for the country. Therefore, the claim appears to be inaccurate ”

Figure 34: RAW-FC dataset sample for label "true"

Claim:

" Police pepper-sprayed a double amputee and removed his prosthetic legs during a June 21, 2020, protest against police brutality in Columbus, Ohio. "

Gold Label: "half"
Predicted Label: "half"

Support Justifications:

"Based on the provided evidence, I conclude that the statement is TRUE. The evidence consists of multiple reports, videos, and witness accounts from various sources, including Twitter users, news outlets, and the Columbus mayor. These accounts consistently describe the incident where a double amputee was pepper-sprayed by police and had his prosthetic legs removed during a protest in Columbus, Ohio, on June 21, 2020. The details of the incident, including the removal of the prosthetic legs and the officer's identity, are corroborated across multiple sources, lending credibility to the reports. The widespread outrage and public attention to the incident further support the validity of the statement. "

Refute Justifications:

" The claim that police pepper-sprayed a double amputee and removed his prosthetic legs during a June 21, 2020, protest against police brutality in Columbus, Ohio, is false. This conclusion is supported by multiple lines of evidence. Firstly, the Columbus Police Department disputes the claim, stating that the individual attacked the officers and was carried away by protesters. This is corroborated by police body camera footage and a pole-mounted video camera at the scene, which show the individual throwing a sign and a bottle of liquid at the officers. Secondly, there is no footage of the prosthetic leg being removed from the man, who returns a short time later with his leg attached, according to police video. This suggests that the prosthetic leg was not taken by the police. Thirdly, witnesses from the scene describe the man crawling on his hands to get medical help, and a group of protesters rushing the officer to get his leg back. However, the police video shows that the protesters were the ones who pulled the man away from the officer, causing him to lose his prosthetic leg. Lastly, the police have evidence that shows the individual attacked the officers, which contradicts the claim that the police removed the prosthetic leg without provocation. In conclusion, the evidence suggests that the claim is false, and the police did not remove the prosthetic leg from the double amputee. The incident was a result of a violent clash between the police and protesters, and the man's prosthetic leg was lost due to the actions of the protesters, not the police. "

Figure 35: RAW-FC dataset sample for label "half"

Claim:

“ 52 thoroughbred horses whose owner died from COVID-19 are destined for slaughter if adopters are not found for them soon. ”

Gold Label: “false”
Predicted Label: “false”

Support Justifications:

“ The claim is likely true. The provided evidence suggests that there is an urgent need to find new homes for 52 thoroughbred horses due to the death of their owner. This implies that the horses are indeed at risk of being sent to slaughter if adopters are not found soon. The fact that a circulating message is asking recipients to help find new homes for the horses further supports the urgency of the situation. Additionally, the specific number of horses mentioned (52) and the breed (thoroughbred) provide context and credibility to the claim. Overall, the evidence suggests that the claim is based on a genuine concern for the welfare of the horses and is not a fabricated or exaggerated story ”

Refute Justifications:

“ The claim that 52 thoroughbred horses are destined for slaughter if adopters are not found for them soon is challenged by several key factors. Firstly, the claim has been circulating for several years, with the original post dating back to 2011. According to credible reports on horse-related forums, all 52 horses were actually rehomed, with most going to family friends of the deceased owner. This suggests that the claim is not based on current events, but rather a recycled and outdated story. Furthermore, the claim has undergone subtle changes over the years, including the reason for the horses being in danger. Initially, the post stated that the horses would be sent to a glue factory, but more recently, the reason cited is the owner' 's passing due to COVID-19. However, there is no evidence to suggest that the horses are currently in danger or that they are being considered for slaughter. In fact, according to TheHorse. Com, all 52 horses were able to find homes within a week of the initial post in 2011. This contradicts the claim that the horses are in imminent danger of being slaughtered. Therefore, based on the available evidence, it appears that the claim is not supported by facts and is likely a recycled and outdated story ”

Figure 36: RAW-FC dataset sample for label "false"

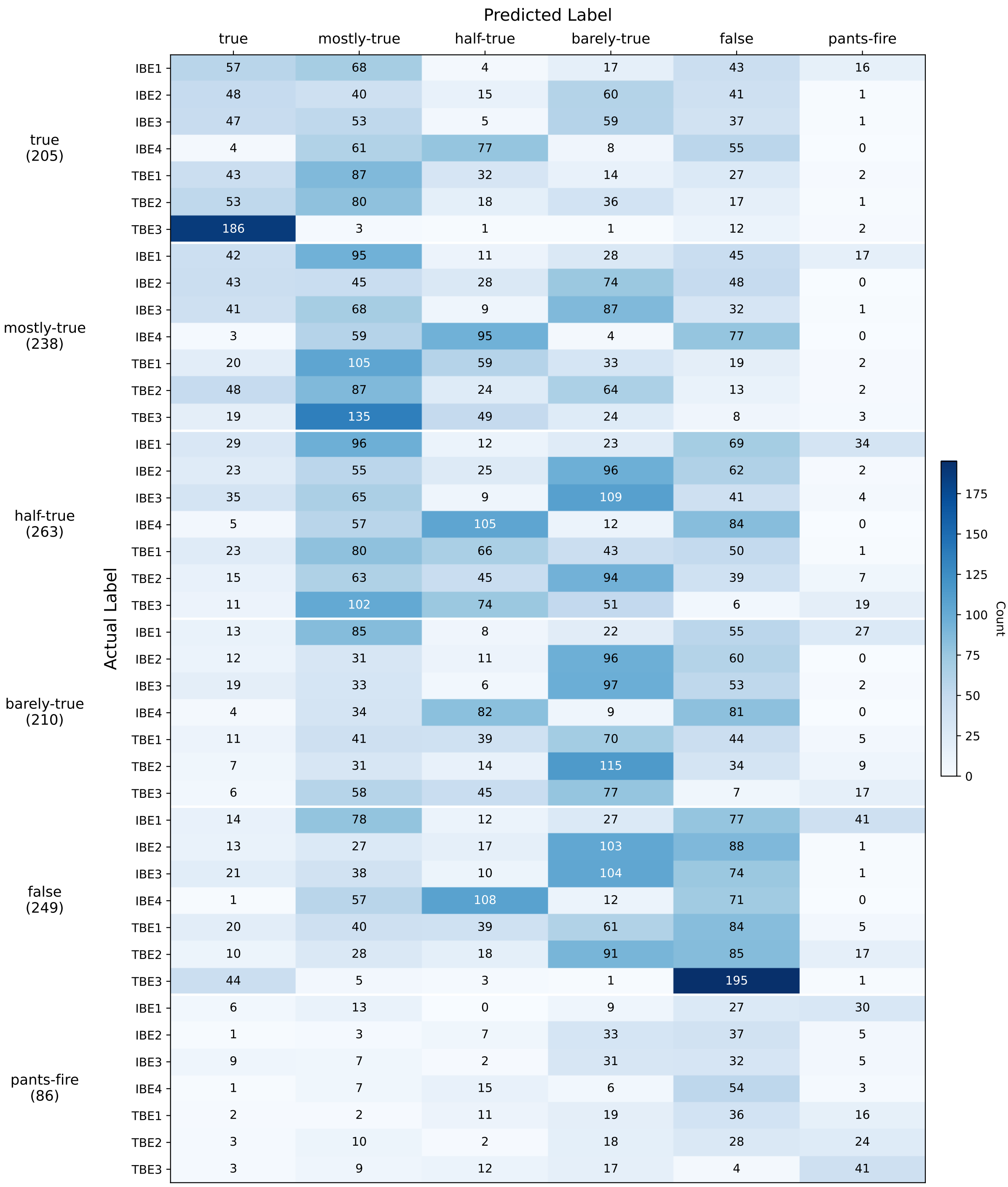


Figure 37: Confusion matrix for LIAR-RAW dataset for IBEs and TBEs.

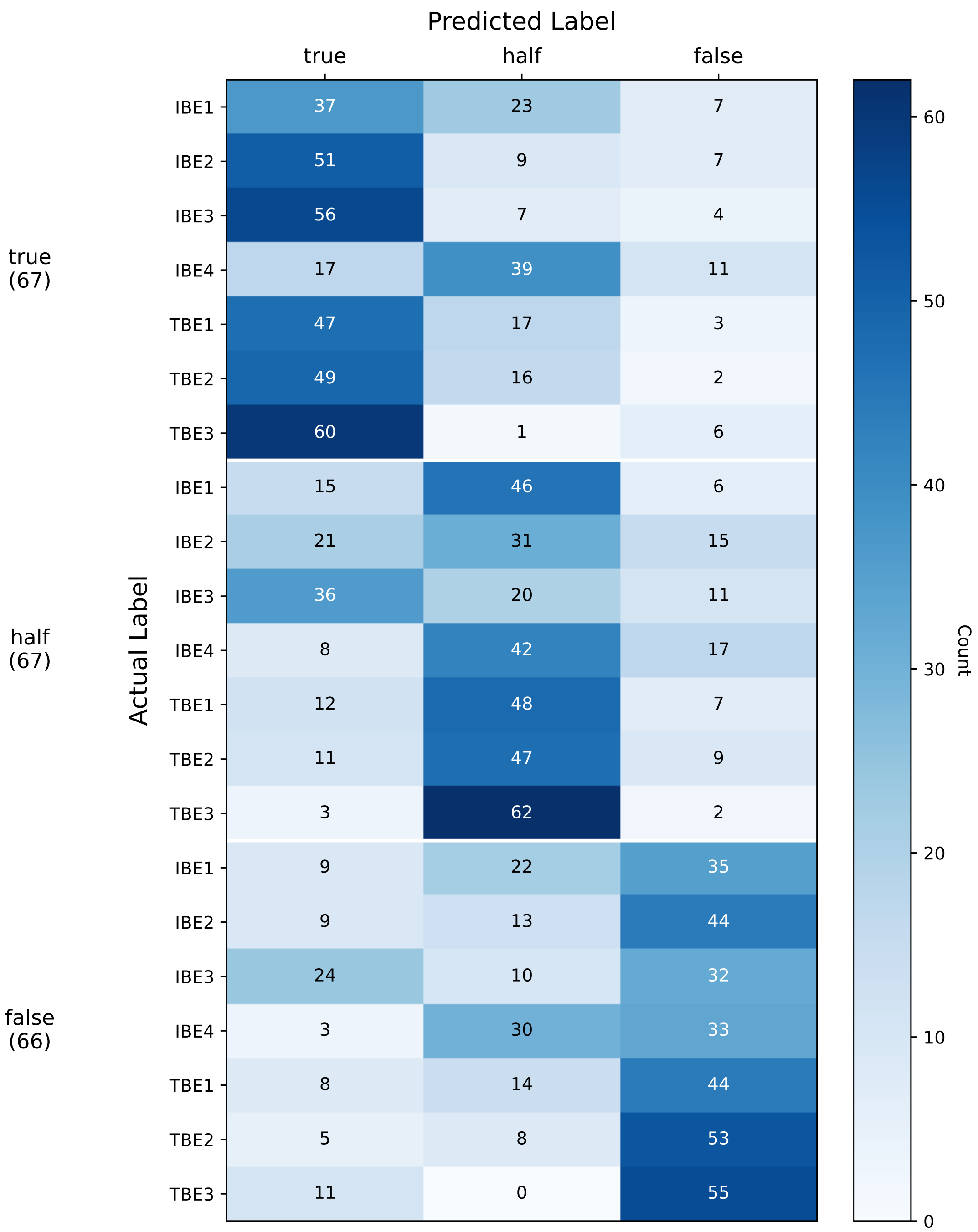


Figure 38: Confusion matrix for RAW-FC dataset for IBEs and TBEs.