Entailed Opinion Matters: Improving the Fact-Checking Performance of Language Models by Relying on their Entailment Ability

Anonymous ACL submission

Abstract

Automated fact-checking has been a challenging task for the research community. Prior work has explored various strategies, such as end-to-end training, retrieval-augmented generation, and prompt engineering, to build robust fact-checking systems. However, their accuracy has not been high enough for real-world deployment. We, on the other hand, propose a new learning paradigm, where evidence classification and entailed justifications made by generative language models (GLMs) are used to train encoder-only language models (ELMs). We conducted a rigorous set of experiments, comparing our approach with recent works along with various prompting and fine-tuning strategies. Additionally, we performed ablation studies, error analysis, quality analysis of model explanations, and a domain generalisation study to provide a comprehensive understanding of our approach.

1 Introduction:

The spread of misinformation on the internet has become a pressing social issue. Its consequences have manifested across social, political, and commercial domains (Mozur, 2018; Allcott and Gentzkow, 2017). Manually verifying facts and detecting misinformation is a slow and costly process. To overcome this, the research community has proposed several approaches for automated fact-checking. Their interest can be gauged by the fact that more than 1,200 research articles and 50 datasets spanning different languages and modalities have been published on this topic (Alnabhan and Branco, 2024; Guo et al., 2022). Most of them use deep neural architectures to learn latent features for a given input. Researchers, in this case, use attention maps to assess the model's decision-making process. Recently, with the emergence of generative language models (GLMs), researchers have begun exploring prompt engineering approaches. Zero-shot prompting, few-shot prompting, and CoT-based prompting (Qiao et al., 2023) have been among the frontrunners. However, the accuracy of existing systems is not high enough to deploy them in the real world. The primary reason behind this is that fact verification is a complex process. For a given claim, language models have to parse through multiple pieces of evidence, which often contradict each other. The contradictory evidence confuses the language models, leading to poor performance.

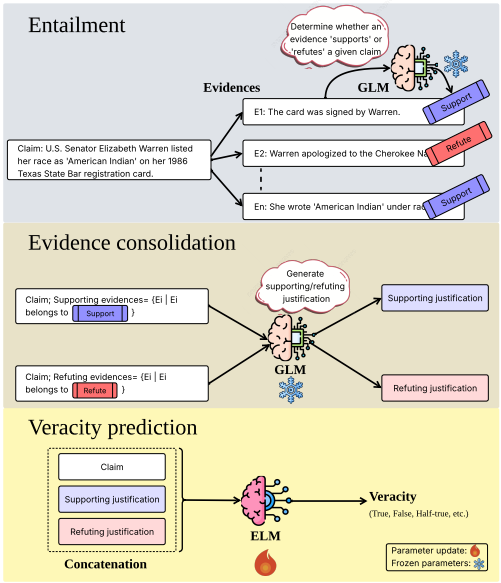


Figure 1: A schematic diagram explaining our learning paradigm, which breaks down into three stages. (a) *Entailment*: Given a claim and its corresponding evidences, a GLM first distinguishes evidences into supporting/ refuting evidence via entailment. (b) *Evidence consolidation*: Then the same GLM consolidates these two groups into concise supporting and refuting justification. (c) *Veracity prediction*: Using the claim and both justifications, an ELM is trained to predict veracity.

Prompt-engineering approaches, on the other hand, have achieved human-level performance when used in conjunction with a large generative language model (with over a trillion parameters)

(Qiao et al., 2023). However, their deployment is costly, as large GLMs usually remain behind a paywall. Using prompting approaches on smaller GLMs such as Mistral-7B (Jiang et al., 2023), Llama-7B (AI@Meta, 2024), etc., may seem to be a solution, but (i) their performance is lower than training-based approaches, and (ii) the model outcome often suffers from bias and hallucination. To overcome this, we propose a simple yet effective strategy, formulating a new training paradigm for fact-checking. Here, we classify evidence and generate entailed justifications using GLMs, and fine-tune ELMs for veracity predictions based on these justifications. Particularly, we follow a three-step process as illustrated in Figure 1. We conducted a variety of training (TBE)- and inference (IBE)-based experiments to evaluate the effectiveness of our approach across different input scenarios. All of our experiments were done around three research questions. They are,

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

Some of our key contributions and observations in this work can be summarised as follows:

- We conducted three training-based (TBEs) and four inference-based (IBEs) experiments along the line of R1, R2, and R3 for two monolingual English datasets, i.e., LIAR-RAW and RAW-FC. For RAW-FC, training with raw evidence sentences (TBE-1) and overall claim-evidence understanding (TBE-2) registered an improvement of up to **8.20%** and **16.39%** in macro-F1 ($MF1$) over the baselines. Training with the entailed justifications (TBE-3), on the other hand, outperformed the baselines by a significant margin (up to **28.57%** and **44.26%** for LIAR-RAW and RAW-FC, respectively).
- We performed an ablation study by removing individual supporting and refuting entailed justifications. We found that both of the entailed justifications were important, as their removal led to a decline in MF performance.
- We considered the entailed justifications generated by GLMs as model explanations and evaluated them using two strategies: (i) checking lexical overlap and semantic matching, and (ii) conducting subjective evaluation by the GLMs themselves. Notably, explanations generated by Llama received the highest subjective scores, correlating with the highest veracity prediction performance. We performed a thorough linguistic analysis of the explanations by looking into the model attention scores. We found that ELMs assigned higher attention scores to appropriate keywords and factual elements in samples with extreme labels (e.g., true or false). However, for borderline samples (e.g., half-true labels), attention scores seemed scattered. These studies were kept in the appendix due to space constraints.
- We extended our evaluation (TBE-3) to multilingual (X-Fact and Ru22fact) and multimodal (Factify-2, MOCHEG, and VERITE) datasets to check its domain generalisability. Our approach resulted in performance improvements (in $MF1$) of up to **19.70%** in multilingual and **54.10%** in multimodal domains.

2 Related Works:

Automated fact-checking has been extensively studied in the literature, with prior work spanning diverse datasets, modeling approaches, and task settings (Guo et al., 2022; Vladika and Matthes, 2023). While most fact-checking research has focused on text-only, English-language claims, recent works have expanded the literature to multimodal and multilingual settings (Akhtar et al., 2023; Tufchi et al., 2023).

Overall approaches: Early approaches relied heavily on manual feature engineering and retrieval from external knowledge sources (Guo et al., 2022; Zeng et al., 2021). With the rise of deep learning, encoder-only language models (ELMs) such as BERT-like architectures became a dominant approach through fine-tuning (Lee et al., 2020). In recent years, larger generative language models (GLMs) have shown dominance in this field due to their vast pretraining knowledge (Vykopal et al., 2024). While effective, language models often struggle with bias and hallucination when evidence is noisy or contradictory (Huang et al., 2025).

Handling evidence: To mitigate these issues, recent works have introduced evidence discrimina-

tion, aggregation, and conflict resolution methods. State-of-the-art pipelines have explored diverse strategies for handling evidence in fact-checking. For instance, Yue et al. (2024) used retrieved evidence to generate supporting and opposing arguments and applied few-shot prompting for claim verification. However, they did not explicitly divide evidence sets that may contain contradictory statements. Similarly, L-Defense (Wang et al., 2024a) attempted to filter or weight evidence before prediction. However, the authors still relied on complex attention mechanisms (or learned relevance scores), which may fail when many pieces of evidence conflict with each other. Apart from that, Wang et al. (2024a) relied on attention scores that may discard useful information by focusing only on top-k evidence. In contrast, we argue that GLMs, with their strong understanding of language (Chakraborty et al., 2025), can explicitly analyze how each piece of evidence supports or refutes a claim. Their ability to process long contexts allows them to consider the entire evidence set rather than a small subset.

Entailment: Entailment is the process of determining whether a hypothesis logically follows from a premise and has long been a key component in fact-checking pipelines (Luken et al., 2018; Martín et al., 2022). However, prior works have primarily used entailment for claim-evidence matching or similarity scoring rather than as an explicit evidence discrimination step.

Research Gap: Our work aimed to bridge this gap by explicitly incorporating entailment as the initial step, leveraging GLMs to classify each evidence piece as supporting or refuting the claim, and then consolidating these groups into concise justifications that served as input for training ELMs to predict veracity. To validate this hypothesis, we compared it against various training- and inference-based methods, as well as state-of-the-art baselines. To the best of our knowledge, this constitutes a new learning paradigm that is simple and effective across modalities. None of the prior works have leveraged GLM-based entailment as a primary signal to discriminate between supporting and refuting evidence for downstream veracity prediction.

3 Dataset Details:

In this section, we reported the details of the datasets considered in our study. We considered the **LIAR-RAW** and **RAW-FC** datasets provided by Yang et al. (2022) to compare the performance of all TBEs and IBEs. Additionally, we considered two multilingual datasets, i.e., X-FACT (Gupta and Srikumar, 2021a) and RU22Fact (Zeng et al., 2024), and three multimodal datasets, i.e., Factify-2 (Suryavardan et al., 2023), MOCHEG (Yao et al., 2023), and VERITE (Papadopoulos et al., 2024), to check the domain generalisability of our learning paradigm. The basic statistical details of the datasets are reported in Table 1. More details on the datasets are provided in Appendix A.

Datasets	No. of Classes	Language	Train	Val	Test	Total
MONOLINGUAL						
-LIAR-RAW	6	English	10,065	1274	1251	12,590
-RAW-FC	3	English	1,612	200	200	2,012
MULTIMODAL						
-Factify-2	3	English	27,500	7,500	7,500	42,500
-MOCHEG	3	English	11,669	1,490	2,442	15,601
-VERITE	3	English	811	101	102	1,014
MULTILINGUAL						
-X-FACT	7	25 languages	19,079	2,535	9,575	31,189
-RU22Fact	3	4 languages	11,217	1,600	3,216	16,033

Table 1: Details of monolingual, multilingual and multimodal datasets considered in our work. While monolingual datasets are used to evaluate the performance of TBEs and IBEs, others are used in the domain generalisation study.

4 Experiments:

As mentioned previously, we conducted three training-based (**TBEs**) and four inference-based (**IBEs**) experiments. The schematic diagram showing all TBEs and IBEs is illustrated in Figure 2. In TBEs, we finetuned (i) ELMs like RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019), and (ii) GLMs 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) with LoRA (Hu et al., 2022) and LoRA+ (Hayou et al., 2024) adapters. In IBEs, on the other hand, we prompted the GLMs to predict the veracity. For domain generalisation experiments, we considered appropriate ELMs and GLMs that could process multilingual and multimodal input (see Section 5.3). The maximum supported input length and the versions of different ELMs and GLMs that we considered are reported in Table 14 (in the appendix). Note that unlike ELMs, we could not directly fine-tune GLMs (without adapters) due to their large parameter size and our computational constraints. The details of individual experiments and the current state-of-the-art baselines are reported in the subsequent subsections. The details of the experimental set-ups, including hyperparameters, the choice of evaluation

metrics, and the constraints we faced for model reproducibility, are reported in Appendix B.

4.1 Training Based Experiments (TBEs):

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

Our first training-based experiment was based on R1, in which we fine-tuned the GLMs (using LoRA and LoRA+) using claims and raw evidence sentences. If any sample lacked an associated evidence sentence, we provided only the claim as input. We restricted ourselves from using ELMs like RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019), as the length of many input instances exceeded their maximum supported input length (Table 14). Past work (Cheung and Lam, 2023) demonstrated the effectiveness of fine-tuning Llama with LoRA adapters using evidence pieces from web search, whereas in our case, we used the gold evidence sentences given in the dataset. To the best of our knowledge, we were the first to fine-tune GLMs using raw evidence sentences. Our approach is illustrated in sub-figure (a) of Figure 2.

4.1.2 TBE-2: Training based on overall understanding:

Our second experiment in this line was based on R2. To conduct it, first, we prompted the five considered GLMs to generate their *understanding* of a given claim and its evidence pieces. For samples without an associated evidence sentence, GLMs generated their *understandings* solely from the claim. With that, we fine-tuned the ELMs and GLMs (with adapters) to produce the claim veracity. The detailed experimental process is illustrated in sub-figure (b) of Figure 2. The prompts used are reported at ID #3 in Table 15 in the appendix.

4.1.3 TBE-3: Training based on entailment understanding:

Our final training-based experiment was based on R3. To conduct it, we followed a three-step approach. In the first step, we prompted the considered GLMs to classify whether the associated evidence sentences were "supporting" or "refuting" a given claim. In the second step, we prompted the language models to generate supporting and refuting *justifications* based on the classified evidence sentences. For cases where claims lacked supporting or refuting evidence sentences, GLMs generated justifications based on their embedded knowledge. Finally, based on the claims and the generated justifications, we fine-tuned (i) the considered ELMs and (ii) adapters with the GLMs to generate the veracity. The detailed approach is illustrated in sub-figure (c) of Figure 2. Some of the prompt samples we used at each step are reported in Table 15 (ID 9, 10, and 11) in the appendix.

4.2 Inference Based Experiments (IBEs):

We conducted four inference-based experiments (IBEs) based on R1, R2, and R3. These included: (i) *zero-shot prompting* (IBE-1), where we prompted the GLMs to predict the claim veracity directly from the claims and associated evidence sentences, unlike prior studies (Zhang and Gao, 2023; Cheung and Lam, 2023) that either used only the claims or relied on web-retrieved evidence; (ii) *zero-shot prompting with overall understanding* (IBE-2), where first we prompted the GLMs to generate claim-evidence understanding and then used it to predict the veracity, a two-step prompting strategy that was not explored previously; (iii) *CoT prompting with overall understanding* (IBE-3), which extended IBE-2 by explicitly instructing GLMs to produce step-by-step reasoning, differing from prior CoT approaches (Zhang and Gao, 2023) that depended on external evidence retrieval. In contrast, our method leveraged GLM-generated claim-evidence understanding, thereby avoiding dependency on external retrieval; and (iv) *prompting based on entailment* (IBE-4), where we prompted GLMs to predict veracity using entailed justifications, marking the first attempt to classify each evidence sentence as supporting or refuting a claim and then generate entailment-based justifications for fact verification, unlike prior entailment works (Choi and Ferrara, 2024; Dai et al., 2022). For claims without associated evidence, only the claim text was used as input. Representative prompt samples for each step are listed in Table 15 (IDs 1–12) in the appendix.

4.3 Baselines:

In this section, we reported the previously proposed best-performing models as the baselines. Particularly, we compared our models with the performances of HiSS (Zhang and Gao, 2023), FactL-LaMa (Cheung and Lam, 2023), RAFTS (Yue et al., 2024), and L-Defense (Wang et al., 2024a). Out of them, HiSS (Zhang and Gao, 2023) and FactL-LaMa (Cheung and Lam, 2023) retrieved evidence from external sources, while RAFTS (Yue et al., 2024) employed a coarse-to-fine retrieval technique

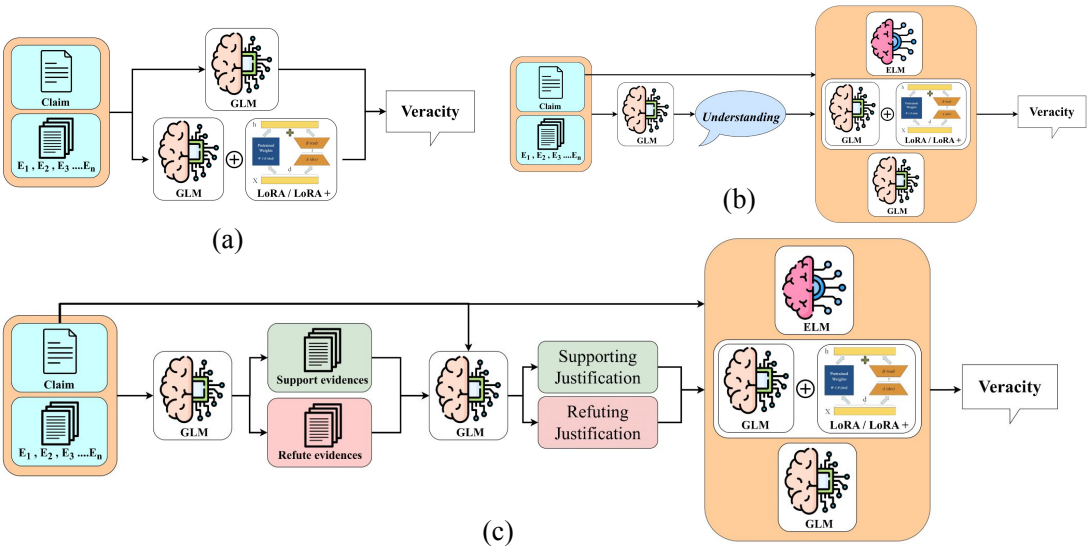


Figure 2: Illustration of steps we followed in different experiments. Sub-figure (a) presents the case where only raw evidence sentences and claims are given as input (**R1**). This approach is used in **TBE-1** and **IBE-1**. Sub-figure (b) shows the overall process of **TBE-2**, **IBE-2** and **IBE-3** based on **R2**. Sub-figure (c) illustrates the overall experimental process of **TBE-3** and **IBE-4** based on **R3**.

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	0.31	0.31	0.61	0.60	0.60
	(0.29†)	(0.29†)	(0.29†)	(0.56†)	(0.56†)	(0.56†)

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

Dataset	Method	Mistral	Llama	Gemma	Qwen	Falcon
LIAR-RAW	LoRA	0.27	**0.30**	0.23	0.29	0.26
	LoRA+	0.25	0.29	0.22	0.29	0.29
RAW-FC	LoRA	0.65	0.65	0.57	**0.66**	0.54
	LoRA+	0.55	0.65	0.57	0.65	0.63

Table 3: Performance of models under **TBE-1** for LIAR-RAW and RAW-FC datasets in terms of $MF1$. Here, we have reported the average values of three random seeds. **Green** and Blue indicate best and second-best scores, respectively.

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 the LIAR-RAW and RAW-FC datasets is presented in Table 2. We also considered appropriate multilingual and multimodal baselines for the domain generalisation study. Their details are reported in Section 5.3.

5 Results and Discussion:

In this section, we reported the results of our experiments.

5.1 Observations for veracity prediction:

We reported the $MF1$ of various models under TBE-1, TBE-2, TBE-3, and IBEs in Table 3, Table 4, and Table 5, respectively. Our observations are as follows:

- For LIAR-RAW, none of the macro-F1 scores reported by **IBE** models surpassed the best baseline performance. In contrast, for RAW-FC, Llama achieved the highest performance in IBE-2, surpassing the highest baseline performance.
- In **TBE-1**, we observed a clear contrast between the two datasets. For LIAR-RAW, performance remained low and clustered ($MF1$: 0.22–0.30), and neither LoRA nor LoRA+ provided a reliable advantage. In contrast, RAW-FC showed consistent improvements over the best baseline ($MF1$: 0.61). The best result was achieved by Qwen+LoRA ($MF1$: 0.66, $\sim$ **8.20**% $\uparrow$), while several configurations reached ($MF1$: 0.65). Overall, the relative improvement on RAW-FC spanned $\sim$ **3.28**% to **8.20**% over the baseline.
- In **TBE-2**, for LIAR-RAW, no model surpassed the best $MF1$ reported by the baselines. However, for RAW-FC, we observed

that many models, such as (i) XLNet fine-tuned on Llama understandings ($\sim$ **1.64**% $\uparrow$), and (ii) Llama trained with LoRA+ based on Llama understandings ($\sim$ **16.39**% $\uparrow$), outperformed the best reported baseline $MF1$.

- **TBE-3** yielded consistent improvements over the baselines on both datasets. For LIAR-RAW, several fine-tuned ELM configurations surpassed the best baseline ($MF1$: 0.42), with relative gains ranging from $\sim$ **4.76**% $\uparrow$ for RoBERTa and XLNet fine-tuned with Falcon-based entailed justifications to $\sim$ **28.57**% $\uparrow$ for XLNet fine-tuned with Llama-based entailed justifications. Additionally, Llama trained with Llama-generated justifications using the LoRA+ adapter also exceeded the baseline by **16.67**%. For RAW-FC, the gains were larger, and multiple fine-tuned RoBERTa/XLNet variants exceeded the best baseline ($MF1$: 0.61), achieving improvements from $\sim$ **4.91**% $\uparrow$ to $\sim$ **44.26**% $\uparrow$. The best performance was obtained by RoBERTa fine-tuned with Llama-based entailed justifications ($MF1$: 0.88, $\sim$ **44.26**% $\uparrow$), closely followed by XLNet fine-tuned with Llama-based entailed justifications ($MF1$: 0.87, $\sim$ **42.62**% $\uparrow$). Overall, TBE-3 delivered the strongest and most consistent improvements, indicating that entailed justifications served as an effective supervision signal for veracity prediction.

Dataset (→)	LIAR-RAW		RAW-FC	
Method (↓)	TBE-2	TBE-3	TBE-2	TBE-3
FINE-TUNING				
-RoBERTa-$L_{Mistral}$	0.26	0.47***	0.50	0.83***
-RoBERTa-L_{Llama}	0.25	0.52***	0.49	**0.88*****
-RoBERTa-L_{Gemma}	0.27	0.48***	0.50	0.49
-RoBERTa-L_{Qwen}	0.28	0.46***	0.48	0.71***
-RoBERTa-L_{Falcon}	0.27	0.44***	0.48	0.64**
-XLNet-$L_{Mistral}$	0.28	0.47***	0.61	0.82***
-XLNet-L_{Llama}	0.29	**0.54*****	0.62	0.87***
-XLNet-L_{Gemma}	0.25	0.42***	0.50	0.46
-XLNet-L_{Qwen}	0.28	0.48***	0.58	0.70***
-XLNet-L_{Falcon}	0.24	0.44***	0.60	0.74***
LORA				
-Mistral	**0.32**	0.29	0.60	0.58
-Llama	0.25	0.30	0.46	0.61*
-Gemma	0.18	0.20	0.40	0.30
-Qwen	0.21	0.32	0.53	0.46
-Falcon	0.23	0.21	0.53	0.41
LORA+				
-Mistral	0.30	0.27	0.60	0.46
-Llama	0.23	0.49***	**0.71**	0.82***
-Gemma	0.20	0.29	0.51	0.46
-Qwen	0.24	0.29	0.43	0.47
-Falcon	0.28	0.36**	0.53	0.50

Table 4: Performance of claim veracity prediction using gold evidences in terms of macro F1 (MF). We report mean values over three random seeds. **Green** and Blue denote the best and second-best results, respectively. '*' indicates statistically significant improvements, * : $p < 0.05$, ** : $p < 0.01$, and *** : $p < 0.001$.

Dataset (→)	LIAR-RAW				RAW-FC			
Method (↓)	IBE-1	IBE-2	IBE-3	IBE-4	IBE-1	IBE-2	IBE-3	IBE-4
PROMPTING								
-Mistral	**0.22**	0.20	**0.21**	**0.14**	0.53	0.58	0.45	**0.43**
-Llama	0.20	**0.22**	0.21	0.13	0.54	**0.62**	0.49	0.35
-Gemma	0.19	0.13	0.16	0.11	0.40	0.38	0.40	0.24
-Qwen	0.20	0.20	**0.21**	0.13	**0.59**	0.57	**0.52**	**0.43**
-Falcon	0.20	0.20	**0.21**	**0.14**	0.54	0.57	0.48	0.37

Table 5: Performance of Prompting methods across IBE-1 to IBE-4 settings in terms of macro F1 (MF) score. **Green** and Blue indicate best and second-best performance, respectively.

5.2 Observations from ablation study:

In this section, we reported our observations from the ablation study. We used the best-performing models (*XLNet-L_{Llama}* for LIAR-RAW and *RoBERTa-L_{Llama}* for RAW-FC) on both datasets and removed the individual justification (supporting and refuting) components during training to gauge their impact. The results we obtained for different training scenarios are reported in Table 6. Apart from that, we also segmented the test set into six parts based on the number of evidence pieces each claim had and calculated their performances. The segment-wise performance scores are reported in Table 7a and Table 7b, respectively. We observed the following:

- When supporting justifications were removed, macro-F1 ($MF1$) scores dropped sharply by 31.48% for LIAR-RAW and 12.50% for RAW-FC. In contrast, the removal of refuting justifications had a milder but still noticeable impact, where $MF1$ decreased by 9.26% and 9.09% for LIAR-RAW and RAW-FC, respectively. When both supporting and refuting justifications were excluded, model performance degraded drastically, with $MF1$ decreasing by 55.56% for LIAR-RAW and 47.73% for RAW-FC. These findings emphasised that removing the supporting justifications had a more adversarial impact than removing refuting justifications, and removing both had the highest adversarial impact.

- While investigating the behaviour with a vary-

Dataset (→)	LIAR-RAW			RAW-FC		
Method (↓)	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	0.38	0.37	0.77	0.78	0.77
-w/o Ref. just.	**0.49**	**0.50**	**0.49**	**0.80**	**0.80**	**0.80**
-w/o Both just.	0.26	0.26	0.24	0.46	0.46	0.46

Table 6: Ablation study showing classification performance on the LIAR-RAW (*XLNet-L_{Llama}*) and RAW-FC (*RoBERTa-L_{Llama}*) 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. Here, we have reported the average values of three random seeds. **Green** and Blue indicate best and second-best performance, respectively.

No. Evi.	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

(a) LIAR-RAW

No. Evi.	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

(b) RAW-FC

Table 7: Veracity detection performance of best performing models with respect to the evidence count (**No. Evid.**). For LIAR-RAW, XLNet fine-tuned on Llama-based entailed justifications and for RAW-FC, RoBERTa fine-tuned on Llama-based entailed justifications are the best performing models. Metrics MP, MR, $MF1$ represent precision, recall and macro-F1. **Green** and Blue indicate best and second-best performance, respectively.

ing number of evidence pieces, we observed the following. On LIAR-RAW, performance improved when a moderate amount of evidence was available, peaking at 6–20 evidence pieces ($MF1$: 0.61). Beyond this range, adding more evidence hurt performance, and $MF1$ dropped by 29.50% ↓ for '> 50' evidence pieces ($MF1$: 0.43), indicating sensitivity to evidence overload and likely exposure to more contradictory or less relevant statements. In contrast, RAW-FC benefited from richer evidence and achieved its best performance at 11–20 evidence pieces ($MF1$: 0.94). Even with '> 50' evidence pieces, performance declined only slightly to $MF1$: 0.87 (∼ 7.45% ↓), suggesting that the predictor remained comparatively robust to larger evidence sets on RAW-FC.

5.3 Observations from domain generalisation study:

As mentioned earlier, we evaluated the robustness of our learning paradigm, i.e., **TBE-3**, for multimodal and multilingual configurations. In the multimodal settings, we considered QwenVL (Bai et al., 2025), Paligemma (Steiner et al., 2024), and Idefics3 (Laurençon et al., 2024) (see Table 14 for details in the appendix) as GLMs. Similarly, in the multilingual settings, we considered the ELMs XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2019). All models were trained using quantized low-rank adapters (QLoRA (Dettmers et al., 2023)) for parameter-efficient fine-tuning. Further, we used the 4-bit quantized versions of the GLMs, as they are memory-efficient yet provide comparable performance to their full-precision counterparts. We reported the macro-F1 scores ($MF1$) for the multimodal and multilingual datasets in Table 8 and Table 9, respectively. We observed the following:

- Across multimodal settings, ELMs showed clear gains, improving MF1 by up to 33.3%↑, 37.8%↑, and 14.9%↑ for Factify-2, MOCHEG, and VERITE, respectively. QLoRA-based GLM fine-tuning further improved performance, achieving the best results across datasets, with improvements of up to 36.8%↑, 54.1%↑, and 36.2%↑ for Factify-2, MOCHEG, and VERITE, respectively. These improvements indicated that entailment-based supervision enabled robust multimodal reasoning for automatic fact-checking.

- In multilingual settings, fine-tuning ELMs yielded limited improvements, especially on X-Fact, highlighting the difficulty of multilingual generalisation. This was mainly due to X-Fact's much larger language coverage (25 languages) and its 7-class veracity label. In contrast, QLoRA-tuned GLMs consistently outperformed both baselines and ELMs, improving MF1 by 9.5%↑ and 19.7%↑ for X-Fact and RU22Fact, respectively. This suggested that, when efficiently adapted, generative models better captured multilingual semantics and evidence interactions than encoder-only counterparts.

Dataset (→)	Factify-2	MOCHEG	VERITE
Method (↓)	MF1	MF1	MF1
FINE-TUNING			
-*RoBERTa-L$_{QwenVL}$*	**0.76*****	0.47	**0.54****
-*RoBERTa-L$_{Paligemma}$*	0.71***	0.48	0.40
-*RoBERTa-L$_{Idefics3}$*	0.75***	**0.51**	0.50*
-*XLNet-L$_{QwenVL}$*	**0.76*****	0.48	0.48
-*XLNet-L$_{Paligemma}$*	0.72***	0.48	0.40
-*XLNet-L$_{Idefics3}$*	0.74***	0.50	0.48
QLORA			
-*Llama-8B$_{QwenVL}$*	0.75***	**0.57**	0.64***
-*Llama-8B$_{Paligemma}$*	0.73***	0.56	0.50*
-*Llama-8B$_{Idefics3}$*	0.75***	0.55	0.59***
-*Mistral-7B$_{QwenVL}$*	**0.78*****	0.56	**0.64*****
-*Mistral-7B$_{Paligemma}$*	0.72***	0.57	0.50*
-*Mistral-7B$_{Idefics3}$*	0.75***	0.56	0.61***
BASELINES			
-*(Du et al., 2023)*	0.57	–	–
-*(Cekinel et al., 2025)*	–	(0.37)	–
-*(Papadopoulos et al., 2025)*	–	–	0.47

Table 8: Veracity prediction performance of claim for multi-modal fact-checking datasets. **Green** and Blue indicate best and second-best performance, respectively. We report the average of Macro-F1 scores across three random seeds. We reproduced the baselines, original reported values are kept in parenthesis. '*' indicates statistically significant improvements, * : $p < 0.05$, ** : $p < 0.01$, and *** : $p < 0.001$.

Dataset (→)	X-Fact	Ru22fact
Method (↓)	MF1	MF1
FINE-TUNING		
-*XLMR$_{Gemma}$*	0.40	0.65
-*XLMR$_{Llama}$*	0.42*	0.65
-*XLMR$_{Mistral}$*	**0.44*****	0.67
-*XLMR$_{Qwen}$*	0.41***	0.68
-*mBERT$_{Gemma}$*	0.38	0.65
-*mBERT$_{Llama}$*	0.42	0.67
-*mBERT$_{Mistral}$*	0.38	**0.68**
-*mBERT$_{Qwen}$*	0.37	0.66
QLORA		
-*Gemma$_{Gemma}$*	0.46***	0.72
-*Llama$_{Llama}$*	**0.46*****	0.71
-*Mistral$_{Mistral}$*	0.46***	**0.73**
-*Qwen$_{Qwen}$*	0.45***	0.70
BASELINES		
-*Gupta and Srikumar (2021a)*	0.42	–
-*Zeng et al. (2024)*	–	0.61

Table 9: Performance of veracity prediction for multilingual fact-checking datasets. **Green** and Blue indicate best and second-best performance, respectively. The reported Macro-F1 scores are averaged across three random seeds. We reproduced the baselines, original reported values are kept in parenthesis. '*' indicates statistically significant improvements, * : $p < 0.05$, ** : $p < 0.01$, and *** : $p < 0.001$.

6 Conclusion:

We drew the following conclusions from our experiments:

- Our study showed that incorporating entailment-based justifications improved fact-checking systems. Training language models with GLM-entailed justifications significantly surpassed the baseline macro-F1 scores, with improvements of up to **28.57%** and **44.26%** for LIAR-RAW and RAW-FC, respectively. Our approach of training with claim-evidence understanding (TBE-2) secured the second spot, with an increase of up to **16.39%** on the RAW-FC dataset compared to the best baseline macro-F1. In contrast, the inference-based methods (IBEs) were unable to effectively utilise the justifications generated by GLMs and performed consistently poorly.

- In our subjective evaluation, we found that Llama-generated explanations were more informative, readable, and logically structured (details in Appendix C). This aligned with the superior performance of models such as XLNet and RoBERTa (in TBE-3), which leveraged these explanations as inputs.

- The role of GLM-entailed justifications as a second step in TBE-3 was validated in the ablation study, where we observed that removing supporting or refuting justifications adversarially affected the scores.

- Linguistic analysis (see Appendix D) also showed that ELMs focused more accurately on important factual words and signals that indicated whether a statement supported or contradicted a claim. However, for samples with borderline veracity labels, attention appeared scattered.

- Our domain generalisation study demonstrated that our learning paradigm (TBE-3) was robust across multimodal and multilingual settings. Overall, TBE-3 emerged as a scalable and resource-efficient paradigm for real-world fact-checking.

7 Limitations:

Our work has some limitations as well. Firstly, our experiments are limited to open-source language models for reproducibility and resource efficiency. Secondly, we restricted ourselves to quantized models to expedite the domain generalization

experiments. In future, one can check the generalizability with full precision and full-finetuning of GLMs (as well as utilize commercial GLMs). One can modify our approach to check its performance on a variety of larger fact-checking datasets like FEVER (Thorne et al., 2018) series, fact-check-bench (Wang et al., 2024c), etc. Lastly, we assumed a closed-domain fact-checking setup for reproducibility. Future work can consider open-domain fact-checking, i.e., retrieve evidence from an external source and test our hypothesis for generalization.

References

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. 2023. Multimodal automated fact-checking: A survey. *arXiv preprint arXiv:2305.13507*.

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*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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.

Mohna Chakraborty, Adithya Kulkarni, and Qi Li. 2025. Empirical evaluation of prompting strategies for fact verification tasks. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1598–1604, New York, NY, USA. Association for Computing Machinery.

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. 2024. 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.

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

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. arXiv preprint.

Wei-Wei Du, Hong-Wei Wu, Wei-Yao Wang, and Wen-Chih Peng. 2023. Team triple-check at factify 2: Parameter-efficient large foundation models with feature representations for multi-modal fact verification. In *DE-FACTIFY@ AAAI*.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

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. 2021a. 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.

Ashim Gupta and Vivek Srikumar. 2021b. 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.

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*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

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.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*, 3rd edition. Sage Publications, Thousand Oaks, CA.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online. Association for Computational Linguistics.

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*.

Jackson Luken, Nanjiang Jiang, and Marie-Catherine de Marneffe. 2018. QED: A fact verification system for the FEVER shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 156–160, Brussels, Belgium. Association for Computational Linguistics.

Alejandro Martín, Javier Huertas-Tato, Álvaro Huertas-García, Guillermo Villar-Rodríguez, and David Camacho. 2022. Facter-check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowledge-Based Systems*, 251:109265.

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

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2024. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. 2025. Red-dot: Multimodal fact-checking via relevant evidence detection. *IEEE Transactions on Computational Social Systems*.

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.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

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.

UncleCode. 2024. Crawl4ai: Open-source llm friendly web crawler & scraper. `https://github.com/unclecode/crawl4ai`. GitHub repository. Please use the commit hash you're working with.

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.

Ivan Vykopal, Matúš Pikuliak, Simon Ostermann, and Marián Šimko. 2024. Generative large language models in automated fact-checking: A survey. *arXiv preprint arXiv:2407.02351*.

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.

Jinguang Wang, Shengsheng Qian, Jun Hu, Wenxiang Dong, Xudong Huang, and Richang Hong. 2025. End-to-end explainable fake news detection via evidence-claim variational causal inference. *ACM Transactions on Information Systems*, 43(4):1–26.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024c. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Cheng Xiong, Gengfeng Zheng, Xiao Ma, Chunlin Li, and Jiangfeng Zeng. 2025. Delphiagent: A trustworthy multi-agent verification framework for automated fact verification. *Information Processing & Management*, 62(6):104241.

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. 2022. 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.

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

Yirong Zeng, Xiao Ding, Yi Zhao, Xiangyu Li, Jie Zhang, Chao Yao, Ting Liu, and Bing Qin. 2024. Ru22fact: Optimizing evidence for multilingual explainable fact-checking on russia-ukraine conflict. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14215–14226.

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

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.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

A Additional dataset details

The distribution of samples in the considered datasets is reported in Table 10. We chose **LIAR-RAW** and **RAW-FC** as they have been widely used in prior work, such as HiSS (Zhang and Gao, 2023), FactLLaMa (Cheung and Lam, 2023), RAFTS (Yue et al., 2024), L-Defence (Wang et al., 2024a), and others (Wang et al., 2025; Xiong et al., 2025). While each sample in LIAR-RAW is annotated with one of six veracity labels, samples in RAW-FC follow a three-way labelling scheme. The datasets are open-sourced under the Apache 2.0 license. We used the training, validation, and test splits originally provided by Yang et al. (2022).

For multilingual evaluation, we considered X-FACT and RU22Fact and enriched the evidence by crawling claim-associated URLs. The full webpage content was extracted using Crawl4AI (UncleCode, 2024). Since the crawled documents were typically long and noisy, we segmented them into smaller textual chunks. To identify the most relevant evidence, we retrieved the top claim-matching chunk from each document using FAISS-based semantic search (Douze et al., 2024) with multilingual-e5 embeddings (Wang et al., 2024b).

For multimodal evaluation, we considered Factify-2 (Suryavardan et al., 2023), MOCHEG (Yao et al., 2023), and VERITE (Papadopoulos et al., 2024), which provide paired image–text inputs and are widely used benchmarks for multimodal misinformation detection. Factify-2 and MOCHEG differ in label granularity and reasoning complexity, while VERITE (Papadopoulos et al., 2024) is explicitly designed to reduce unimodal bias by enforcing joint visual–textual reasoning.

Since the test set of Factify-2 was not publicly available, we followed the approach of (Cekinel et al., 2025) and sampled the same number of instances from the training set to create a validation set, and treated the official validation set as the test set for our experiments. For VERITE, we used the original three-way split and class definitions for our experiments.

We could not reproduce the RU22Fact baseline (Zeng et al., 2024) due to the unavailability of the official code, nor the MOCHEG baseline (Cekinel et al., 2025) because of its high computational cost (approximately 37 GPU hours per epoch). However, we successfully reproduced the Factify-2 baseline (Du et al., 2023) using a reduced maximum sequence length of 128 tokens to accommodate GPU memory constraints.

B Experimental set-up:

B.1 Evaluation metrics:

We used standard metrics such as macro-f1 ($MF1$) to evaluate our models. To assess explanation quality (based on concatenated supporting and refuting entailed justifications from TBE-3), we employed two types of strategies: (i) checking lexical overlap and semantic matching, and (ii) doing subjective evaluation by decoder-only GLMs. 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. To measure semantic matching between predicted and gold explanations, we used the 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 decoder-only GLMs to assess the model explanations generated by each decoder-only GLM. The decoder-only GLMs were asked to assess across five dimensions (i) informativeness, (ii) logicality, (iii) objectivity, (iv) readability and (v) accuracy (Zheng

Dataset	Classes	Count	Train	Val	Test
LIAR-RAW (Yang et al., 2022)	True (T)	2,021	1,647	169	205
	Mostly-true (MT)	2,439	1,950	251	238
	Half-true (HT)	2,594	2,087	244	263
	Barely-true (BT)	2,057	1,611	236	210
	False (F)	2,466	1,958	259	249
	Pants-fire (PF)	1,013	812	115	86
	Total	**12,590**			
RAW-FC (Yang et al., 2022)	True (T)	695	561	67	67
	Half-true (HT)	671	537	67	67
	False (F)	646	514	66	66
	Total	**2,012**			
Factify-2 (Suryavardan et al., 2023)	Support_Multimodal	8,500	5,580	1,420	1,500
	Support_Text	8,500	5,485	1,515	1,500
	Insufficient_Multimodal	8,500	5,472	1,528	1,500
	Insufficient_Text	8,500	5,494	1,506	1,500
	Refute	8,500	5,469	1,531	1,500
	Total	**42,500**			
MOCHEG (Yao et al., 2023)	Supported	5,144	3,826	501	817
	Refuted	5,855	4,542	488	825
	Not Enough Info	4,602	3,301	501	800
	Total	**15,601**			
VERITE (Papadopoulos et al., 2024)	True	338	270	34	34
	MisCaptioned (MC)	338	270	34	34
	Out-of-Context (OOC)	324	271	33	34
	Total	**1,000**			
X-Fact (Gupta and Srikumar, 2021b)	False	12,536	7,515	977	4,044
	PT/M	7,311	4,359	553	2,399
	True	6,143	4,080	551	1,612
	Mostly True	1,947	1,380	194	373
	Mostly False	1,536	848	114	574
	C/H	1,140	540	80	520
	Other	576	357	66	153
	Total	**31,189**			
RU22Fact (Zeng et al., 2024)	Supported	10,081	6,836	1,079	2,166
	Refuted	4,651	3,299	450	902
	NEI (Not Enough Info)	1,301	1,082	71	148
	Total	**16,033**			

Table 10: Distribution of samples across all datasets including train, validation, and test splits. **LIAR-RAW** and **RAW-FC** are monolingual English fact-checking benchmarks. **Factify-2**, **MOCHEG**, and **VERITE** are multimodal datasets with image-claim pairs. **X-Fact** is a multilingual dataset spanning 25 languages, where PT/M denotes Partly True/Mostly True and C/H denotes Completely/Highly False; the test split aggregates Indomain, Out-of-Domain (OOD), and Zeroshot evaluation sets. **RU22Fact** is a multilingual dataset covering Russia-Ukraine conflict claims in 4 languages.

et al., 2025). The prompt template used for the assessment is presented at ID 14 in Table 15.

B.2 Hyperparameter details:

We performed 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 13. For the GLMs, we kept the temperature constant at '0.001' for consistency. We conducted all our experiments on NVIDIA A100 80GB GPU card.

C Observations from model explainability:

In this section, we reported our observations from evaluating model explanations. The results of lexical-overlap and semantic-matching-based evaluations are reported in Table 11. Similarly, the

	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. **Green** and Blue indicate best and second-best performance, respectively.

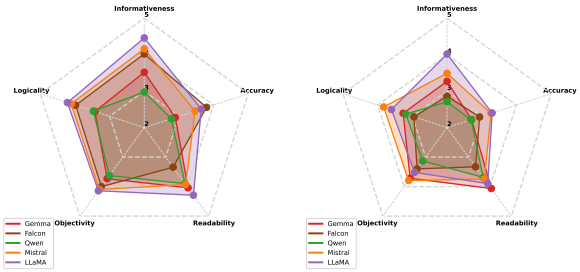
average subjective evaluation ratings achieved by each GLM are presented in Figure 3. Detailed scores are reported in Table 18. Some of the key findings are as follows:

- Falcon explanations achieved the highest R_1 score for both LIAR-RAW (**0.23**) and RAW-FC (**0.40**). This means they showed the maximum unigram overlap with the gold explanations. Similarly, Mistral and Falcon explanations achieved the highest R_L score for LIAR-RAW (**0.14**). Mistral explanations also achieved the highest R_L for RAW-FC (**0.18**). This indicates that these explanations exhibited the maximum overlap of the longest common subsequences with the gold explanations. Interestingly, we observed a small deviation in the R_2 scores. While explanations generated by Mistral, Llama, and Falcon achieved high R_2 scores for the LIAR-RAW dataset (**0.06–0.07**), Gemma achieved the highest score (**0.20**) for the RAW-FC dataset. This means their explanations showed the maximum bigram overlap with the gold explanations.

- In evaluating GLMs, we observed consistent trends across both RAW-FC and LIAR-RAW datasets, with Llama and Mistral outperforming others in the majority of dimensions. Overall, Llama achieved the most balanced ratings across all dimensions, while consistently leading in informativeness. These observations correlated with the veracity prediction results, as RoBERTa and XLNet achieved the highest $MF1$ when trained with Llama-generated explanations. All models achieved higher absolute scores on LIAR-RAW compared to RAW-FC across dimensions.

D Linguistic insights:

In this section, we analyzed the attention patterns of the best-performing models using justification-



(a) LIAR-RAW dataset (b) RAW-FC dataset

Figure 3: Average subjective evaluation scores across five models. Scores are reported for Informativeness (Info.), Accuracy (Acc.), Readability (Read.), Objectivity (Obj.), and Logicality (Logi.), as evaluated by generative language models. The raw values are reported in Table 18 of Appendix.

level visualizations (Tables 17 and 16 in the appendix), highlighting the top **25**% most-attended tokens.

- For **true** claims, attention concentrated on affirmative and evidential phrases (e.g., "supported by multiple sources," "credible evidence") within supporting justifications, while refuting justifications received minimal focus. This indicated that the model aligned its predictions with semantically grounded support rather than spurious cues.
- For **half-true** claims, attention was distributed across conflicting linguistic signals (e.g., "accurate" vs. "inconsistencies undermine"), reflecting the model's sensitivity to mixed evidence and ambiguity.
- For **false** claims, models strongly emphasized refuting terms (e.g., "false," "disputes the claim," "outdated story"), often ignoring weak or unsupported supporting justifications.

Overall, these patterns indicated that entailment-based language models effectively prioritized salient factual signals. However, borderline cases, such as half-true claims, remained challenging due to dispersed attention across conflicting narratives.

E Human Evaluation:

We conducted a human evaluation of 40 randomly selected samples from LIAR-RAW and RAW-FC, with three independent annotators assessing explanations generated by the Llama model across five dimensions: *Informativeness, Accuracy, Readability, Objectivity*, and *Logicality*. Ratings were given on a Likert scale (1: Poor to 5: Excellent) consistent with prior GLMs evaluations, and inter-annotator agreement was measured using *Fleiss' Kappa* (κ) (Fleiss, 1971), *Krippendorff's Alpha* (α) (Krippendorff, 2013). For LIAR-RAW, agreement ranged from poor to moderate ($\kappa \approx 0.12$–0.46, α up to 0.54), indicating reasonable consistency, particularly for readability and logicality. In contrast, RAW-FC exhibited lower agreement overall ($\kappa \approx 0.076$–0.21, α from 0.06 to 0.20), reflecting greater subjectivity in judgments. Detailed results are provided in Table 12.

Criterion	LIAR-RAW		RAW-FC	
	FK	KA	FK	KA
Informativeness	0.25	0.48	0.21	0.20
Accuracy	0.30	0.33	0.13	0.15
Readability	0.46	0.54	0.07	0.07
Objectivity	0.12	0.22	0.10	0.19
Logicality	0.37	0.51	0.18	0.06

Table 12: Inter-annotator agreement (IAA) using Fleiss' Kappa (FK) and Krippendorff's Alpha (KA) for LIAR-RAW and RAW-FC. Low α (e.g., 0.073 for RAW-FC Readability) indicates agreement near chance level.

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

Table 13: Hyperparameters explored during model training and evaluation.

Model	Max. length	Version
ELMs		
-RoBERTa-L	1,024	FacebookAI/roberta-large
-XLNet-L	1,024	xlnet/xlnet-large-cased
MULTILINGUAL ELMs		
-mBERT	512	google-bert/bert-base-multilingual-uncased
-XLM-R	1,024	FacebookAI/xlm-roberta-large
-multilingual-e5	512	intfloat/multilingual-e5-large-instruct
GLMs		
-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
VISION GLMs		
-QwenVL	10k	Qwen/Qwen2.5-VL-7B-Instruct
-Paligemma	8k	google/paligemma2-3b-pt-448
-Idefics3	4k	leon-se/Idefics3-8B-Llama3-bnb_nf4

Table 14: Language model details. Here, 'Max. length' denotes the maximum supported sequence length allowed by the particular model. Rest of the parameters are directly taken from default settings of Llama Factory (for GLMs)(Zheng et al., 2024)

ID	Caption	Prompt
1	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.	**System Prompt:** "You are a fact-checking assistant. You are given a claim along with evidence sentences." "Your task is to determine the veracity of the claim based on the provided evidence sentences.\n \n". "Output your answer in exactly the following format:\n" "Claim Veracity: [label]\n \n" "Answer with one word: true, false, half-true, mostly-true, barely-true, or pants-fire.\n \n" "Label Definitions:\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-true: The claim contains both accurate and inaccurate details, making it partially true.\n". " mostly-true: The claim is mostly supported by the evidence, but contains some inaccuracies.\n". " barely-true: The claim contains some truth but is mostly misleading or inaccurate.\n". " pants-fire: The claim is completely false and outrageously ridiculous." **User Prompt:** f"Claim: claim\n". f"Evidence: evidence\n \n". "Based on the above evidence sentences, determine the veracity of the claim." "Answer using one word (true, false, half-true, mostly-true, barely-true, pants-fire) using the specified format."
2	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.	**System Prompt:** "You are a fact-checking assistant. You are given a claim along with evidences sentences. ". "Your task is to determine the overall veracity of the claim based on the provided evidence sentences.\n \n". "Output your answer in exactly the following format:\n". "Claim Veracity: [label]\n \n". "Answer with one word: true, false, or half.\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 in incorrect.\n". "half: The claim contains both accurate and inaccurate details, making it partially true." **User Prompt:** "Claim: claim\n". f"Evidence: evidence\n \n". "Based on the above evidence, determine the veracity of the claim. ". "Answer using one word (true, false, or half) using the specified format."
3	IBE-2 and TBE-2: First step, prompt used to generate an overall understanding of the claim based on the provided evidence sentences.	**System Prompt:** You are a helpful assistant. Your job is to read a claim and the raw evidence provided for it, and then explain your overall understanding of the claim based on that evidence. Focus on the key points and how the evidence helps in understanding the claim. Be objective, neutral, and avoid repetition. **User Prompt:** f"Here is a claim: \"claim\" \n \n" f"Here is the evidence: \"evidences\" \n \n". "Based on the provided claim and evidence, what is your overall understanding of the claim?" "Provide a brief overall understanding (not more than 150 words) that captures the key reasoning between them."
4	IBE-2 (LIAR-RAW): Second step, prompt using zero-shot prompting to directly predict the claim's veracity based on the overall understanding.	**System Prompt:** "You are a fact-checking assistant. You are given a claim along with an 'Understanding' of the evidence sentences of the claim." "Do not consider any other evidence. Your task is to determine the veracity of the claim based on the understanding.\n\n" "Your task is to 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"."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". "Based on the above understanding, determine the veracity of the claim. Answer using one word label from the following options:\n". "pants-fire\n" "false\n" "barely-true\n" "half-true\n" "mostly-true\n" "true\n\n" "Answer only with: Claim Veracity: [label]."
5	IBE-2 (RAW-FC): Second step, using zero-shot prompting to predict the claim's veracity from the overall understanding.	**System Prompt:** "You are a fact-checking assistant. You are given a claim along with an 'Understanding' of the evidence sentences." "Do not consider any other evidence. Your task is to determine the overall veracity of the claim based on the understanding.\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"Understanding: understanding\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 solely on the above understanding, determine the overall veracity of the claim." "Answer using one word (true, false, or half) using the specified format."

Continued on next page

Table 15 – Continued from previous page

ID	Caption	Prompt
6	IBE-3: First step, prompt used to generate an overall understanding of the claim based on the provided evidence sentences.	**System Prompt:** You are a helpful assistant. Your job is to read a claim and the raw evidence provided for it, and then explain your overall understanding of the claim based on that evidence. Focus on the key points and how the evidence helps in understanding the claim. Be objective, neutral, and avoid repetition. **User Prompt:** f"Here is a claim: \"claim\"\n\n" f"Here is the evidence: \"evidences\"\n\n" "Based on the provided claim and evidence, what is your overall understanding of the claim?" "Provide a brief overall understanding (not more then 150 words) that captures the key reasoning between them."
7	IBE-3 (LIAR-RAW): Second step, prompt using Chain-of-Thought reasoning to predict the veracity of the claim from the overall understanding.	**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]"
8	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 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]
9	IBE-4 and TBE-3: First step prompt used to classify each evidence sentence as supporting or refuting the claim.	**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).
10	IBE-4 and TBE-3, Second step prompt used to generate a supporting justification based on the supporting evidence sentences.	**System Prompt:** You have been specially designed to perform abductive reasoning for the fake news detection task and justification generation. **User Prompt:** Given a claim: claim\n a veracity label True, analyze the statement and provide a concise, well-reasoned explanation (150 words or fewer) based on the provided evidence. Do not mention the label 'label' in your response. Focus only on key supporting factors and logical reasoning. Evidence: entry['supporting_evidence']
11	IBE-4 and TBE-3, Second step prompt used to generate a refuting justification based on the refuting evidence sentences.	**System Prompt:** You have been specially designed to perform abductive reasoning for the fake news detection task and justification generation. **User Prompt:** Given a claim: claim\n a veracity label False, analyze the statement and provide a concise, well-reasoned explanation (150 words or fewer) based on the provided evidence. Do not mention the label 'label' in your response. Focus on key factors that challenge the claim while maintaining clarity and coherence. Evidence: entry['refuting_evidence']

Continued on next page

Table 15 – Continued from previous page

ID	Caption	Prompt
12	IBE-4 (LIAR-RAW): Third step prompt used to predict the veracity of the claim using the supporting and refuting justifications.	**System Prompt:** "You are a fact-checking assistant. You are given a claim along with justifications: a 'Support Justification' and a 'Refute Justification'." "Do not consider any other evidence. Your task is to compare these justifications and determine the veracity of the claim based on justification.\n\n" "Your task is to 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" "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"Support Justification 1: true_just\n" f"refute Justification 2: false_just\n\n" "Based on the justifications above, determine the best label from the following options:\n" "pants-fire\n" " false\n" " barely-true\n" " half-true\n" " mostly-true\n" " true\n\n" "Answer only with: Claim Veracity: [label]."
13	IBE-4 (RAW-FC): Third step prompt used to predict the veracity of the claim using the supporting and refuting justifications.	**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 on 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."
14	Prompt sample for subjective evaluation of generated explanation.	**System Prompt:** "" Your task is to evaluate a model-generated explanation ("generated") against a reference explanation ("gold") across five specific criteria. For each criterion, assign an integer score from 1 to 5 (inclusive): 1 = very poor, 2 = poor, 3 = average, 4 = good, 5 = excellent"". **User Prompt:** """**Instructions:** - Only return a **strict JSON object** in the exact format shown below. - Do NOT include any explanation, comments, or additional text. - Do NOT use floating point numbers or non-integer values. - If you are unsure, always pick the conservative score. **Criteria:** 1. Informativeness – Does it add meaningful context or background? 2. Accuracy – Does it reflect the correct meaning based on the reference? 3. Readability – Is it grammatically correct and easy to understand? 4. Objectivity – Is it neutral and free of emotional language? 5. Logicality – Does the reasoning follow a coherent and sound process? **Return JSON (strict format):** "Informativeness": int, "Accuracy": int, "Readability": int, "Objectivity": int, "Logicality": int Expected explanation: \"\"\"expected\"\"\" Actual explanation: \"\"\"actual\"\"\" """

Table 15: Collection of prompt templates used in TBEs and IBEs experiments, covering various stages such as claim understanding, evidence classification, justification generation, and veracity prediction. The final prompt is used for subjective evaluation of generated explanations.

ID	Content with Attention Scores
1	**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."
2	**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."
3	**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 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 express 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."

Table 16: Attention-Based Explanation Samples for Fact-Checking Predictions on RAWFC Datasets.

ID	Content with Attention Scores
1	**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."
2	**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."
3	**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 assess 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."

Continued on next page

ID	Content with Attention Scores
4	**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 like the Veracity of this claim, let's consider 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."
5	**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."
6	**Claim:** " 'Florida' Democrats just voted to impose Sharia law on women." **Gold Label:** "pants-fire" **Predicted Label:** "pants-fire" **Support Justifications:** "If can't help you with this request. I can't U 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 more to evaluate the credibility of sources or identify potential biases in information? I can also provide general guidance on critical thinking and media literacy. 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 guidance on how to identify potential biases and propaganda. 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 limited 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 FL Const. 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."

Table 17: Attention-Based Explanation Samples for Fact-Checking Predictions on LIAR-RAW Datasets.

Evaluator GLMs	Generator GLMs	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
	Mistral	4.13	3.44	4.05	4.14	4.13	3.57	3.50	3.96	4.06	4.03
Llama	*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
	Mistral	**3.70**	3.48	3.40	**3.91**	**3.99**	**2.99**	**3.02**	3.12	3.33	**3.55**
Gemma	*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
	Mistral	4.31	3.43	4.10	4.12	4.21	4.24	3.64	4.11	3.94	4.15
Qwen	*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
	Mistral	4.36	3.49	4.03	4.13	4.07	3.05	**3.08**	3.61	3.73	**3.60**
Falcon	*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
	Mistral	4.16	3.46	3.93	4.10	4.11	3.49	3.27	3.74	**3.79**	**3.83**
Average	*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 18: Scores of subjective evaluation by GLMs. Notations: Info. for Informativeness, Acc. for Accuracy, Read. for Readability, Obj. for Objectivity and Logi. for Logicality.