MetaSumPerceiver: Multimodal Multi-Document Evidence Summarization for Fact-Checking

Ting-Chih Chen
Virginia Tech
tingchih@vt.edu

Chia-Wei Tang
Virginia Tech
cwtang@vt.edu

Chris Thomas
Virginia Tech
chris@cs.vt.edu

Abstract

Fact-checking real-world claims often requires reviewing multiple multimodal documents to assess a claim's truthfulness, which is a highly laborious and time-consuming task. In this paper, we present a summarization model designed to generate claim-specific summaries useful for fact-checking from multimodal, multi-document datasets. The model takes inputs in the form of documents, images, and a claim, with the objective of assisting in fact-checking tasks. We introduce a dynamic perceiver-based model that can handle inputs from multiple modalities of arbitrary lengths. To train our model, we leverage a novel reinforcement learning-based entailment objective to generate summaries that provide evidence distinguishing between different truthfulness labels. To assess the efficacy of our approach, we conduct experiments on both an existing benchmark and a new dataset of multi-document claims that we contribute. Our approach outperforms the SOTA approach by 4.6% in the claim verification task on the MOCHEG dataset and demonstrates strong performance on our new Multi-News-Fact-Checking dataset.

1 Introduction

Fact-checking claims on social media platforms poses a significant challenge due to the large volume of new claims constantly being posted without sufficient methods for verification (Aïmeur et al., 2023). Research indicates that manually verifying all aspects of a 200-word claim can require up to four hours of dedicated effort (Vladika and Matthes, 2023). Further, despite the exceptional capabilities of large language models (LLMs) in natural language processing tasks, they still generate content with factual errors. Given that LLMs can produce convincing statements and thus influence beliefs, the potential for hallucinations poses a serious risk of misleading users when deployed for fact-checking (Jakesch et al., 2023b,a; Kreps et al., 2022). This concern highlights the possibility of language models becoming new sources of misinformation and disinformation. The proliferation of misinformation and fake news adds to this predicament, making it increasingly difficult to distinguish between reliable and deceptive content (Goldstein et al., 2023; Spitale et al., 2023). Thus, there is an urgent need for tools capable of succinctly summarizing relevant evidence for fact-checkers, i.e., systems that provide a brief yet comprehensive overview of the relevant evidence to facilitate accurate and reliable assessments. Existing research relying on summarization for fact-checking is ineffective because these methods fail to extract evidence from the resources (Das et al., 2023; Ceron and Carrara, 2023; Berlinski et al., 2023).

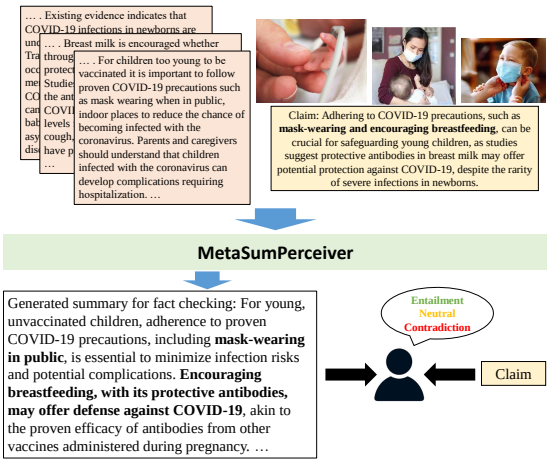

Figure 1: Overview of MetaSumPerceiver (MSP): Using inputs such as documents, images, and claims, MSP generates summaries to facilitate fact-checking. In this example, the summary provides evidence and establishes that the claim in question is entailed by the evidence.

A potential solution to this problem is provided by multimodal summarization (Khullar and Arora, 2020; Liu et al., 2022), which can generate summaries from sources including text, images, videos, and audio. This is a challenging task because each modality might contribute complementary informa-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8742–8757

Table 1: Comparing Multi-News-Fact-Checking with other fact-checking datasets. Topic(s) is the inferred value.

Datasets	#Samples	Source	Topic(s)	Language	Multi-modal	Multi-doc	Verification	Explanations	Text/image retrieval
Zlatkova et al. (2019)	1,233	Snopes, Reuters	< 1,500	English	✔	✗	✔	✗	✗
Cheema et al. (2022)	3,400	X	< 3,400	English	✔	✔	✔	✗	✔
Nielsen and McConville (2022)	12,914	X	26,048	Multi	✔	✔	✔	✗	✗
Yao et al. (2023)	15,601	Politifact, Snopes	< 15,631	English	✔	✗	✔	✔	✔
Nakov et al. (2021)	18,014	X	< 1,312	Multi	✗	✔	✔	✗	✗
Ours	111,905	Multi-News	< 1,500	English	✔	✔	✔	✔	✔

tion, e.g., a bar chart image accompanying relevant facts mentioned in the text. Present methods usually produce summaries given a limited number of inputs (Puduppully and Lapata, 2021; Wang et al., 2022a). The research challenge lies in processing arbitrary inputs from diverse modalities and discerning the explicit relationships between them. However, unlike the standard summarization task, which seeks to summarize the salient content of an article, our objective is to effectively distill claim-specific evidence useful for fact-checking across various modalities.

To train our system to generate summaries useful for human fact-checking, we assess the utility of our summaries at performing entailment (Dagan et al., 2006), a closely aligned task to fact-checking. Our work is orthogonal to prior work in entailment, in that rather than learning to predict the entailment label for the premise-hypothesis pair, we seek to generate the premise for a specific claim from a pool of multimodal data. To address the limitations of applying existing summarization methods for fact-checking, we propose the **MetaSumPerceiver** (MSP) model in Figure 1, where the input consists of a claim, a set of documents and images, and the objective is to generate a summary that expedites the fact-checking process for humans. We initially train the perceiver model with a summarization model. Subsequently, to produce the summary for fact-checking, we employ a proxy reward mechanism to update the summarizer to ensure the generation of an accurate and relevant summary with necessary evidence.

To support research on the task of multi-document fact-checking, we contribute a benchmark (Multi-News-Fact-Checking) of claims and entailment labels whose evidence is drawn from multiple documents. We evaluate our method on the MOCHEG benchmark (Yao et al., 2023) and our new dataset and demonstrate substantial improvements compared to existing baselines. The major contributions of this paper are as follows:

- We present an innovative approach for multi-modal multi-document summarization specifically designed for fact-checking applications.
- We release the Multi-News-Fact-Checking dataset, to support the multimodal multi-document fact-checking summarization task.
- We perform detailed experiments and ablations of our model and loss functions which clearly demonstrate the superiority of our approach over existing methods.

2 Related Work

2.1 Perceiver

The Perceiver architecture (Jaegle et al., 2021) enables scaling transformers to input sequences of arbitrary lengths, by reducing the memory footprint in standard self-attention. Follow-up works, such as Perceiver IO (Jaegle et al., 2022), adapt the original model by presenting a versatile architecture adept at processing data from various settings while ensuring linear scalability with input and output dimensions. The model has demonstrated strong performance on many downstream tasks, including optical flow estimation (Butler et al., 2012) and the GLUE language benchmark (Wang et al., 2018). Our method relies on Jaegle et al. (2022) to process a variable number of arbitrarily long text documents and images. We use the model in sequence with a summarization model to generate a multimodal summary.

2.2 Multimodal Fact-checking Datasets

In the current task of fact-checking using multiple datasets (Zlatkova et al., 2019; Nakov et al., 2021; Cheema et al., 2022; Nielsen and McConville, 2022; Yao et al., 2023), the main sources of data are X and Snopes, with a focus on COVID-19, elections, and the Russo-Ukrainian war. This has led to a couple of issues. First, X has already blocked their API, making it difficult for people to access these datasets. Second, we are seeking a dataset that covers a variety of topics rather than being limited to specific ones. Additionally, we prefer a dataset in the English language. To address these issues, we propose a new multimodal fact-checking

dataset based on Multi-News (Fabbri et al., 2019), which includes information from multiple documents and related images, covering a broader array of topics such as news, policy, weather, sports, etc. Table 1 provides a comparison of the Multi-News-Fact-Checking dataset with the aforementioned datasets.

2.3 Learning From Feedback

Recent advancements in LLMs have revolutionized the AI landscape (Touvron et al., 2023a,b; Driess et al., 2023; OpenAI, 2023). However, because they are mostly trained on data scraped from the web LLMs sometimes produce undesired outcomes, including generating biased or harmful content (Bender et al., 2021). Recognizing the importance of aligning LLMs with human values, has led to efforts in supervised fine-tuning (SFT) with ethical guidelines (Taori et al., 2023). While these efforts demonstrate the potential of integrating human feedback into training using reinforcement learning for user-tailored tasks (Ouyang et al., 2022; Bai et al., 2022), training LLMs to reflect human values is quite challenging. In our work, we adopt the idea of training language models with feedback. However, rather than relying on a human fact-checker, we utilize a surrogate reward model (an entailment model) to stand in the place of a human fact checker, in order to fine-tune the summarizer to generate summaries that give evidence for fact-checking specific claims through Proximal Policy Optimization (PPO) (Schulman et al., 2017; Zheng et al., 2023).

Table 2: Analysis of claims in the Multi-News-Fact-Checking dataset. Top: Entailment accuracy using Llama 2. Bottom: Classification results indicating claim checkworthiness.

Entailment label consistency	**Accuracy (%)**
Entailment claims	78.3
Nentral claims	64.2
Contradiction claims	74.1
Checkworthiness results	**Percentage (%)**
Unimportant factual sentence (UFS)	17.67
Checkworthy factual sentence (CFS)	68.6
Non-factual sentence (NFS)	13.71

3 Multi-News-Fact-Checking Dataset

To train our system, we need a dataset of claims whose facts are drawn from multiple documents along with the entailment label of each claim. We build our dataset on top of the Multi-News summarization dataset (Fabbri et al., 2019), which contains sets of multiple text documents along with human-written summaries of each set. Because the Multi-News dataset lacks claims specifically tailored for fact-checking tasks, we prompt Llama 2 (Touvron et al., 2023b) to generate labeled claims from each set of documents. Within each group of Multi-News documents, we leverage the human-written multi-document summary to generate 30 claims (ten of each entailment type), resulting in a dataset of 1,291,168 labeled claims. The specific prompts contain sections containing a task description, example, and instructions, which are fully detailed in the appendix 9.1. Our dataset contains 111,905 images we obtained by retrieving images from the original articles.

To assess the quality and effectiveness of claim generation, we conducted an evaluation using a scale from 1 to 5, where 1 indicates low quality and 5 indicates high quality. We tested 60 claims and obtained an average score of 3.61 for claim generation quality and effectiveness. In comparison, the claims generated from PRIMERA summaries scored an average of 3.21. To ensure impartiality, the annotators were blinded and asked to rate the claims alongside their respective articles and summaries.

Additionally, in Table 2, we validated claim labels (entailment, neutral, and contradiction) using Llama 2. Specifically, we treated the ground truth label (i.e., the label used to prompt Llama 2 to generate a claim) as the ground truth and prompted Llama 2 in a zero-shot manner to predict the entailment labels. The average accuracy for claim verification was 72.2%, indicating that the generated claims were largely consistently predicted as their intended labels.

To assess the checkworthiness of our generated claims (i.e., to ensure that they are factual claims worth fact-checking), we use a pre-trained model trained on the ClaimBuster dataset (Arslan et al., 2020), as illustrated in Table 2. The model assigns claims into three classes: UFS (unimportant factual claims that are not considered check-worthy), CFS (claims containing factual information of public interest in terms of their veracity), and NFS (sentences that do not contain any factual claims). The result shows that 70% of the prompted claims are check-worthy claims. This outcome substantiates that our prompts are well-designed for this task, and that Llama 2 accurately comprehend the task's intended meaning without misunderstanding.

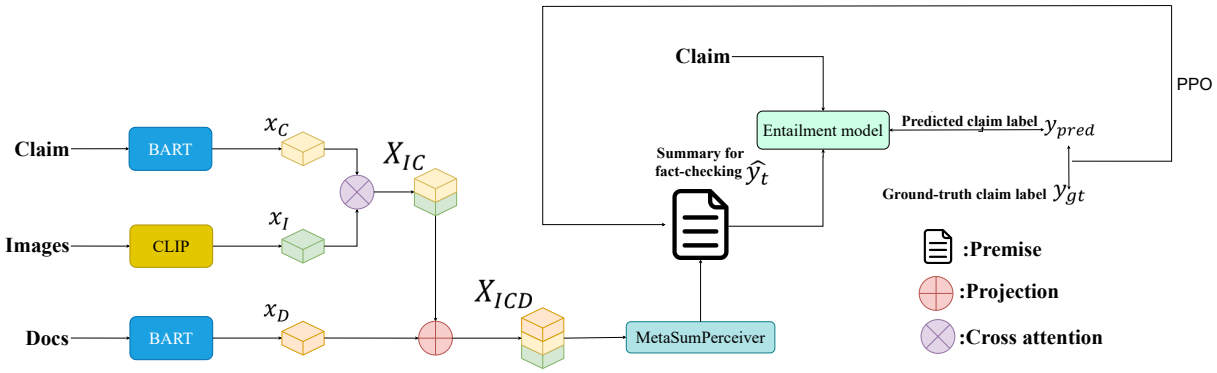

Figure 2: Overview of MetaSumPerceiver (MSP): This figure illustrates the process of generating a summary for fact-checking using MSP, integrating a fixed entailment model for accurate truthfulness labeling. Furthermore, it highlights how PPO is employed to continually refine the summary during the fact-checking process.

4 Approach

In this section, we explain the details of our approach, MSP as illustrated in Figure 2. We also describe the preprocessing steps for both text and image data, the components of our model, and the reinforcement learning methodology we applied to train MSP. Our approach is capable of summarizing multiple multimodal documents consisting of arbitrarily long texts and images. Specifically, we use x_C, x_D, and x_I to represent embeddings for claims, documents, and images, respectively.

4.1 Preprocessing

Due to the sequence length limitation, we utilize a combination of Perceiver with BART (Lewis et al., 2020) and CLIP (Wang et al., 2022b) to extract embeddings. To circumvent exceeding the sequence length, we break down the data into chunks of 1024 tokens. Subsequently, we employ Perceiver to merge these chunks for both textual and visual embeddings. For the textual data, we use BART to obtain text embeddings following (Devlin et al., 2019). As a result, each input text is transformed into a set of token embeddings $x_C \in \mathbb{R}^{n \times D}$ and $x_D \in \mathbb{R}^{m \times D}$, where n and m are the number of tokens and D is the dimension of embedding. Then, we use CLIP (ViT-G-14) to extract visual features for the images. Finally, each input image undergoes a transformation, resulting in a set of visual embeddings. $x_I \in \mathbb{R}^{k \times D}$, where k is the number of tokens and D is the dimension of the embedding.

4.2 Model Training Strategy

Our goal is to generate a textual summary of a set of multimodal documents that enables a fact-checker to determine the veracity of a claim. In order to select relevant visual content from the images, we begin by performing a cross-attention between the images and the claim:

$$X_{IC} = ATTN(Q_{x_C}, K_{x_I}, V_{x_I}) \,, \qquad (1)$$

where the query Q_{x_C} is the claim's sequence of embeddings and K_{x_I} and V_{x_I} are the embedding sequences of visual tokens from the images. We project X_{IC} into the document embedding X_D, which serves as the input for MSP. The output from the cross-attention block, X_{IC}, is initially projected by a linear projection layer with the weight θ. It is then concatenated with x_D, as depicted in the subsequent equation:

$$X_{ICD} = \left[proj(X_{IC}, \theta)^\intercal, X_D^\intercal\right]^\intercal , \qquad (2)$$

where X_{ICD} will be the input to MSP. Prior to training our full model, we pre-train our attention block and summarization model using the Multi-News dataset's human written summaries using the cross-entropy loss function:

$$\mathcal{L}_{\text{sum}} = -\sum_{t=1}^{T} \sum_{i=1}^{N} y_{t_i} \log(\hat{y}_{t_i}), \qquad (3)$$

where T represents the sequence length, N is the vocabulary size, and y_{t_i} and $\hat{y}_{t_i}$ denote the ground truth and predicted probabilities of token i at time step t, respectively. In the remaining text, we omit the summation over the vocabulary for conciseness.

4.3 Fine-tuning The Summarizer

To enhance the summarizer's ability to produce summaries that provide the evidence needed for

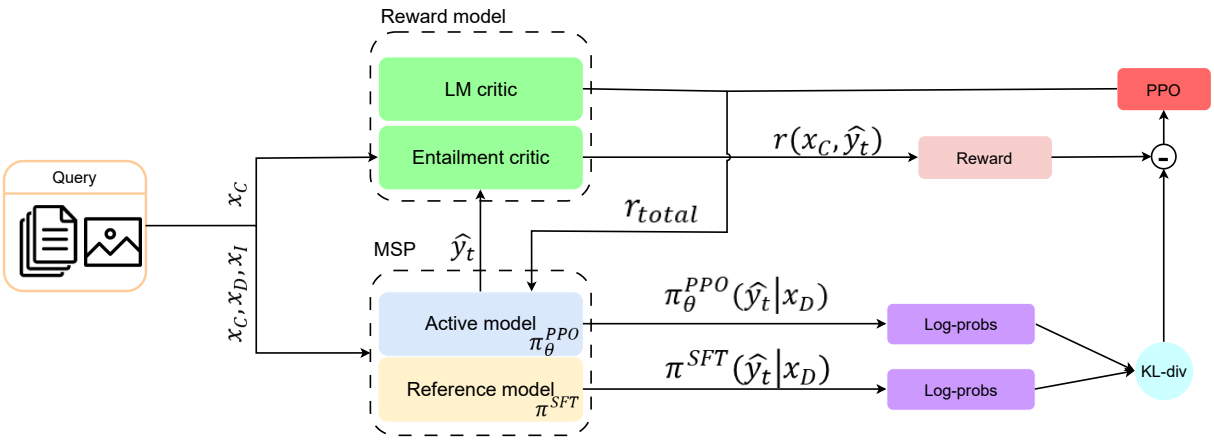

Figure 3: The Proximal Policy Optimization (PPO) process starts with the summarizer generating a response based on the input query. The reward model then assesses this query-response pair, producing a single reward score. Simultaneously, the process calculates the KL-divergence by comparing the likelihood of token sequences in the response with both the currently fine-tuned active model and a pre-trained reference model. The KL-divergence acts as a measure of reward, ensuring that responses from the active model align with those from the reference model. Additionally, we input the summary into Mistral LM to evaluate whether the summary is concise or not. In conclusion, PPO updates the parameters of the active model based on the reward model's output, Mistral LM, and the value of the KL-divergence.

fact-checking claims, we adopt the concept of training a language model using feedback with reinforcement learning. After pretraining the perceiver and summarization models, we employ reinforcement learning with an entailment model serving as a surrogate for a human fact-checker as feedback. We first exclusively apply reinforcement learning to the perceiver. Subsequently, we unfreeze the summarizer and continue training end-to-end with both the perceiver and summarizer. We illustrate our fine-tuning process in Figure 3.

4.3.1 Reward Model For Fact-Checking

Contrary to the approach in reinforcement learning from human feedback, which necessitates a human arbitrator to score the model's outputs, in this study, we train a reward model to act like a human fact-checker to guide the summarizer in producing summaries for fact-checking instead. We utilized a comprehensive dataset consisted with MultiNLI (Williams et al., 2018), Fever-NLI (Thorne et al., 2018), and Adversarial-NLI (ANLI) (Nie et al., 2020), encompassing a total of 763,193 premise-claim pairs. Leveraging this dataset, we fine-tuned DeBERTAV3 (He et al., 2023) for the task of entailment classification using cross-entropy loss. Serving as an entailment classifier, this model achieves accuracy rates of 90.3%, 77.7%, and 57.9% in the MultiNLI, Fever-NLI, and ANLI evaluation datasets, respectively.

4.3.2 Proximal Policy Optimization

We define the score from the reward model as the probability of the ground-truth label given both the claim (as the hypothesis) and the generated summary for fact-checking (as the premise). The formulation for the score from the reward model can be formulated as:

$$r(x_C, \hat{y}_t) = P(y_{gt}|x_C, \hat{y}_t) - 0.5 * \Sigma_{y_{gt} \neq y_{pred}} P(y_{pred}|x_C, \hat{y}_t), \quad (4)$$

where x_C, $\hat{y}_t$, y_{gt} and y_{pred} denote the claim, the generated summary, the groud-truth label of the claim, and the predicted label of the claim, respectively. The value of $P(y_{\{gt,pred\}}|x_C, \hat{y}_t)$ is derived from the trained entailment classifier. The primary objective behind this reward function is to maximize the likelihood that the generated summary for fact-checking contains the facts necessary for the model to predict the claim's ground truth label.

In this paper, we employ PPO as our policy gradient method for reinforcement learning. PPO adds an additional term to the reward function, which imposes a penalty determined by the Kullback-Leibler (KL) divergence between the trained RL policy summarizer, π_ϕ^{PPO}, and the initial supervised summarizer π^{SFT}.

Moreover, we incorporate an extra reward $r_{quality}$, LM critic, to evaluate the quality of the

Table 3: Performance of claim verification in MOCHEG with our method. We separately calculate the precision and recall in supported, refuted, and NEI claim labels. We compare our method with published baselines in Table 4. The labels "Supported," "NEI," and "Refuted" in fact-checking classification are analogous to truthfulness labels. "Supported" aligns with the entailment label, indicating that the hypothesis is similar to the premise. "NEI" corresponds to "not enough information" and is comparable to the neutral label, indicating that the hypothesis includes both information entailed by the premise and information that lacks clarity or confirmation. "Refuted" shares the same classification as contradiction, indicating that the hypothesis is not entailed with the premise.

Setting	Accuracy (%)	Precision (%) Supported	Precision (%) Refuted	Precision (%) NEI	Recall (%) Supported	Recall (%) Refuted	Recall (%) NEI
MSP (Entail. critic) w/ Text Evidence → DeBERTAV3	43.7	79.2	66.9	**33.9**	40.5	30.6	25.8
MSP (Entail. critic) w/ Text + Img Evidence → DeBERTAV3	50.8	83.4	**69.3**	27.3	42.9	34.2	30.9
MSP (Entail. critic) w/ Text Evidence → Llama 2	46.7	80.4	68.1	31.5	37.2	35.4	31.5
MSP (Entail. critic) w/ Text + Img Evidence → Llama 2	**53.7**	**87.3**	60.3	32.4	**48.3**	**36.9**	**34.8**
MSP (Entail., LM critics) w/ Text Evidence → DeBERTAV3	40.2	77.3	63.4	**45.9**	38.2	35.7	28.4
MSP (Entail., LM critics) w/ Text + Img Evidence → DeBERTAV3	47.8	78.1	**67.5**	38.1	39.5	**37.5**	34.1
MSP (Entail., LM critics) w/ Text Evidence → Llama 2	49.3	81.5	65.2	37.4	39.7	31.5	35.7
MSP (Entail., LM critics) w/ Text + Img Evidence → Llama 2	**55.6**	**88.2**	57.5	39.6	**51.2**	32.4	**37.2**

summary, specifically focusing on clarity and conciseness. We utilize Mistral (Jiang et al., 2023) along with a detailed quality testing prompt provided in the appendix to assess this aspect. The assigned reward ranges from 0 to 1. We integrate $r_{quality}$ into our model update in conjunction with the existing r_{total}. The cumulative reward is described as follows:

$$r_{total} = (r_{quality} + r(x_C, \hat{y}_t) - \eta KL(\pi_\phi^{PPO}(\hat{y}_t|x_D), \pi^{SFT}(\hat{y}_t|x_D)))/2, \quad (5)$$

where η represents the KL reward coefficient, which determines the magnitude of the KL penalty, we set it to 0.2 for our model. This coefficient functions as an entropy boost, enhancing exploration throughout the policy domain and urging the model to engage in a diverse set of actions rather than the one currently considered the best. In addition, it inhibits the policy from rapidly committing to a singular strategy, and this encourages outputs from the RL fine-tuned model to not deviate too far from the original model.

MSP is optimized through PPO based on the policy gradient methods that optimize the policy of the model using gradient ascent. The update rule for the policy gradient is given as:

$$\theta \longleftarrow \theta + \alpha \nabla_\theta J(\theta), \quad (6)$$

where α and J_θ denote the learning rate and the expected return under policy π_θ from the model, respectively.

5 Experiments

5.1 Claim Verification

The goal of our method is to generate a summary from multiple documents and modalities that is useful for fact-checking a claim. In order to assess how useful our method is at this task, we compare the performance of our method on MOCHEG, which presents a benchmark and method for multi-document multimodal fact-checking. Specifically, we employed two *fixed* entailment models, namely DeBERTAV3 (He et al., 2023) and Llama 2 (Touvron et al., 2023b), as our surrogate "human" fact checkers to predict the entailment label of a claim given our generated summary. Importantly, we do not fine-tune these models with our generated summaries to avoid biasing the models toward the linguistic or stylistic patterns of the summaries. This ensures that they do not learn spurious features in the downstream task.

As depicted in Tables 3 and 4, our method exhibits superior performance, achieving a SOTA 48.6 F-score in the MOCHEG dataset. Furthermore, according to Table 3, our method demonstrates strong precision performance for the "Supported" label. We conduct separate tests with two distinct critics. It is noted that the most optimal performance was achieved when deploying both the entailment critic and the LM critic. This outcome indicates that our model is proficient in verifying claim labels through clear and concise summaries.

Table 3 reveals that the best results are achieved when inputs incorporate both textual and image evidence. Perhaps unsurprisingly given its size, the zero-shot Llama 2 entailment surrogate model surpasses DeBERTAV3 in performance. Nevertheless, a notable issue persists, where the surrogate entail-

ment models struggle to accurately deal with NEI claim labels.

Table 4 highlights the superiority of our model compared to MOCHEG. In the case of MOCHEG, truthfulness labels are predicted by averaging a stance representation derived from both textual and image evidence. Furthermore, MOCHEG's classifier relies on fixed thresholds, which may not be optimal for every situation. In contrast, our approach involves generating summaries for fact-checking via reinforcement learning with fixed entailment models and LM critic. Although a difference remains in the result of human vs system prediction performance, our model surpasses the prior state-of-the-art system by 4.6% F-score.

Table 4: Performance of claim verification in MOCHEG. DeBERTaV3 and Llama 2 represent the fixed entailment models. "Gold Evidence" denotes ground truth text and image evidence, while "System Evidence" refers to automatically retrieved text and image evidence. "Human" indicates human evaluation.

Setting	F-score (%)
MSP(Entail. critic) w/ Text Evidence → DeBERTAV3	42.7
MSP(Entail. critic) w/ Text + Img Evidence → DeBERTAV3	45.1
MSP(Entail. critic) w/ Text Evidence → Llama 2	43.9
MSP(Entail. critic) w/ Text + Img Evidence → Llama 2	**48.2**
MSP(Entail., LM critics) w/ Text Evidence → DeBERTAV3	44.1
MSP(Entail., LM critics) w/ Text + Img Evidence → DeBERTAV3	46.1
MSP(Entail., LM critics) w/ Text Evidence → Llama 2	44.6
MSP(Entail., LM critics) w/ Text + Img Evidence → Llama 2	**48.6**
MOCHEG w/ Text Evidence	42.7
MOCHEG w/ Image Evidence	40.9
MOCHEG w/ Text and Image Evidence	**44.0**
Human w/o Evidence	20.0
Human w/ System Evidence	62.0
Human w/ Gold Evidence	**70.0**

5.2 Ablations

Additionally, we conducted ablation experiments for claim verification on our Multi-News-Fact-Checking dataset. A comparative analysis of our method with Llama 2 and other offline summarization models, PRIMERA (Xiao et al., 2022), PEGASUS (Zhang et al., 2020) and T5 large (Raffel et al., 2020), is presented in Tables 5 and 6.

Similar to our results in MOCHEG, Tables 5 and 6 show that our approach, when employing the Llama 2 surrogate entailment model, achieves the best performance. Furthermore, we achieve balanced accuracy in both precision and recall, underscoring our method's ability to clearly differentiate between truthful and untruthful labels without bias in predictions. The results highlight the inability of other summarization models to generate summaries useful for fact-checking, which causes the surrogate model difficulty in accurately assessing the truthfulness labels. Furthermore, it is evident that the LM critic significantly aids the entailment model in verifying claim labels effectively. The LM critic ensures that the summary is more concise and clear while retaining the essential meaning in the summary.

In addition, to assess the degree to which our summaries are merely extractive of the source articles, we employed ROUGE to evaluate our summaries alongside the provided articles. Our summaries received a ROUGE score of 0.53, whereas the human-written summaries in the Multi-news dataset scored 0.62. Upon comparison, we believe our summaries do not simply rephrase the source articles.

Moreover, to determine the fidelity and informativeness of the generated summaries, we conducted a human evaluation using a scale from 1 to 5, with 1 indicating low fidelity/informativeness and 5 indicating high fidelity/informativeness. We tested 60 generated summaries and obtained an average score of 3.77 for fidelity and informativeness. Comparatively, the PRIMERA summaries scored 3.35. To ensure impartiality, we blinded the annotators and asked them to rate the generated summaries alongside their respective articles.

Table 5: Performance of claim verification in Multi-News-Fact-Checking dataset. DeBERTAV3 and Llama 2 serve as the fixed entailment models. "Gold Evidence" denotes ground truth text and image evidence, while "System Evidence" refers to automatically retrieved text and image evidence. "Human" indicates human evaluation.

Setting	F-score (%)
PEGASUS → DeBERTAV3	25.4
PEGASUS → Llama 2	**30.8**
T5 large → DeBERTAV3	28.5
T5 large → Llama 2	**32.7**
PRIMERA → DeBERTAV3	38.2
PRIMERA → Llama 2	**38.3**
MSP(Entail., LM critics) → DeBERTAV3	40.1
MSP(Entail. critic) → Llama 2	41.8
MSP(Entail., LM critics) → Llama 2	**43.7**
Human w/o Evidence	23.0
Human w/ System Evidence	65.0
Human w/ Gold Evidence	**76.0**

We also removed the image component from our model and tested it on the FEVER testing dataset. Using Llama2, our zero-shot entailment prediction accuracy was 62%. Using PRIMERA's summaries, the most competitive baseline for multidocument summarization, we obtained 42% on FEVER. Thus, we continue to outperform on this benchmark. Notably, we did not fine-tune the downstream clas-

Table 6: Performance of claim verification in Multi-News-Fact-Checking dataset. We compare our method with Llama 2, and other offline summarization models.

Setting	Accuracy (%)	Precision (%) Entailment	Precision (%) Contradiction	Precision (%) Neutral	Recall (%) Entailment	Recall (%) Contradiction	Recall (%) Neutral
PEGASUS → DeBERTAV3	33.2	64.2	14.7	21.5	37.3	12.4	11.9
PEGASUS → Llama 2	39.5	37.4	23.1	42.8	27.6	24.3	24.0
T5 large → DeBERTAV3	34.8	62.8	17.5	26.2	33.0	18.5	18.2
T5 large → Llama 2	37.2	40.2	32.8	**48.0**	30.5	26.4	26.8
PRIMERA → DeBERTAV3	35.9	68.2	32.7	23.7	35.8	23.8	45.1
PRIMERA → Llama 2	39.2	43.5	47.2	33.1	47.2	35.5	24.9
MSP(Entail., LM critics) → DeBERTAV3	36.9	**74.3**	29.3	28.3	42.5	23.9	**44.8**
MSP(Entail. critic) → Llama 2	42.6	41.0	**53.7**	34.6	54.8	37.8	29.6
MSP(Entail., LM critics) → Llama 2	**46.0**	49.5	49.3	34.1	**56.4**	**44.7**	28.9

Table 7: Performace of explanation generation. Our system outperforms MOCHEG on equivalent settings. Gold Truthfulness denotes ground truth claim label and System Truthfulness means the predicted claim label.

Setting	ROUGE 1 (%)	ROUGE 2 (%)	ROUGE L (%)	BLEU (%)	BERTScore (%)
MOCHEG w/ Gold Evidence, Gold Truthfulness	**45.5**	**27.3**	**35.4**	**21.8**	**89.0**
MOCHEG w/ Gold Evidence, System Truthfulness	43.8	26.3	34.1	20.8	88.8
MOCHEG w/ System Evidence, Gold Truthfulness	35.5	17.4	**26.0**	**10.9**	87.0
MOCHEG w/ System Evidence, System Truthfulness	33.8	16.5	24.8	10.0	86.9
MSP(Entail., LM critics) w/ System Evidence, Gold Truthfulness	**37.8**	**19.4**	24.6	11.4	**88.1**
MSP(Entail., LM critics) w/ System Evidence, System Truthfulness	35.1	16.3	24.9	10.6	87.5

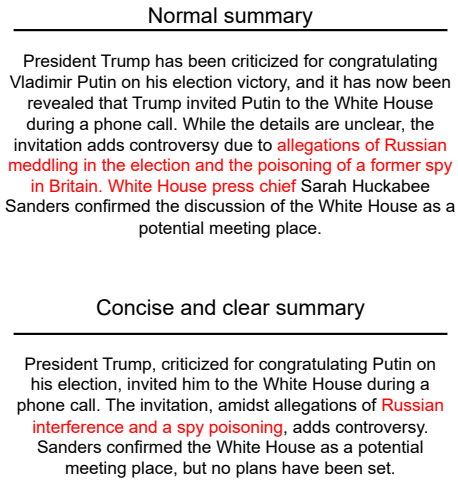

Figure 4: The normal summary is produced by our initial MSP model, while the concise and clear summary is generated using MSP trained with the $r_{quality}$ reward.

sifier on any benchmark, unlike FEVER's evaluation methods that directly train on labeled data in FEVER. Fine-tuning the final classifier would likely yield better results, but this is orthogonal to our contribution. Our zero-shot evaluation on FEVER, compared to the most competitive multi-document summarization baseline, demonstrates that our approach significantly outperformed (62% vs. 42%), underscoring the importance of fact-checking-driven summarization.

5.3 Explanation Generation

In order to assess the degree to which our generated summaries contain the relevant facts necessary to fact check the generated claims, we measure the ability of a method to generate an *explanation* of the predicted truthfulness label using our summary. We adopt a methodology similar to Yao et al. (2023), where we consider the input claim C, its truthfulness label Y_C, and the summary for fact-checking $\{T_1, T_2, ...\}$ generated from MSP. These components are concatenated into an overall sequence X using a separator </s>. During the training of the rationale generator, we employ the actual truthfulness label of each claim as input. Critically, we do not retrain or fine-tune MSP for this task. In the evaluation phase, we utilize the truthfulness label predicted by the fixed entailment models.

Following Yao et al. (2023), we utilize BART to generate the ruling statement. Our evaluation metrics include ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang* et al., 2020). To assess the performance of explanation generation, we compare it with MOCHEG (Yao et al., 2023), as shown in Table 7, Figure 4 and 5.

We observe that our model outperforms MOCHEG's evidence-retrieval based method ("system evidence") on the rationale generation task. In our case, "system evidence" is our generated summary. We note that MOCHEG's method relies on retrieval from a pool of multimodal documents. The ground truth explanations rely on these sentences and thus may share some phrasing. This gives a slight advantage to MOCHEG's method on some metrics that measure n-gram overlap, whereas our method based on summarization may rephrase the same evidence. Nevertheless, we observe that our system outperforms MOCHEG's

Claim	Summary for fact-checking	Image evidence	Truthfulness
The driver in a video appeared to ram a crowd of people, some of whom were identified as Taco Bell employees.	... **The incident, captured on video, shows the driver assaulting an employee before plowing into people outside the restaurant, then crashing through the front.** Two individuals were treated on-site, and one was taken to the hospital. The Charles County Sheriff's Office is investigating, offering a cash reward for information. Charges are pending as they work to identify and locate the driver and occupants.		**Supported**
Criticism over the failure to conduct immediate, widespread coronavirus testing in the U.S. focused on the availability of test kits from the WHO.	... The U.S. faced delays in developing and deploying tests, with defective CDC tests and regulatory issues. **Accusations surfaced that the U.S. rejected WHO's test, but WHO clarified they never discussed providing tests to the U.S.** The CDC's standard protocol involves developing its own tests, and Dr. Anthony Fauci mentioned it would have been nice to have WHO tests as a backup but emphasized the CDC's focus on reliable testing.		**NEI**
A video documents a hot-mic exchange between Fox News reporter John Roberts and a 'fake news tech' in which they admit COVID-19 was a hoax, and that everyone in the news media has already been vaccinated.	... **It does not reveal a bombshell revelation but rather an informal exchange between friends. Mills jokes about the virus being a hoax, and Roberts clarifies that he doesn't think it was a hoax.** The video, not deceptively edited, captures sardonic gallows humor and was shared on social media with misleading claims about the pandemic being a hoax. The conversation mainly revolves around a recent USC study on infection rates in California.		**Refuted**

Figure 5: Explanation generation examples of Multimodal Fact-Checking. The Truthfulness column shows gold labels.

generated explanations.

We further observe that our explanations generated using system evidence and system truthfulness outperform MOCHEG's method, which relies on the ground truth truthfulness label on the BERTScore metric. Overall, these results demonstrate that our summarizer, which was not trained for the rationale prediction task, is capturing relevant evidence across modalities in a short summary better than MOCHEG's evidence retrieval-based approach.

6 Conclusion

We present MetaSumPerceiver (MSP), a summarization model crafted to generate concise and informative summaries specifically tailored for fact-checking claims in intricate multimodal datasets. The model's adaptable architecture can handle varying numbers of documents and input types, encompassing documents, images, and claims, by leveraging a perceiver-based design. We train our model using a RL approach, aiming to produce summaries that are instrumental in verifying the accuracy of claims. Moreover, our reward function is designed to generate more concise and clear summaries, aiding in the verification of diverse claims with the assistance of the LM critic. Our experimental assessments on the MOCHEG and our Multi-News-Fact-Checking datasets highlight MSP's robust performance in claim verification and explanation generation tasks and demonstrate its effectiveness in real-world fact-checking scenarios. This contribution underscores MSP's potential to streamline fact-checking processes in today's multimodal information landscape. Finally, we release the publicly accessible Multi-News-Fact-Checking dataset, aimed at assisting researchers in developing multi-document fact-checking methods.

7 Acknowledgements

We acknowledge Advanced Research Computing at Virginia Tech for providing computational resources and technical support that have contributed to the results reported within this paper. We also thank all reviewers for their comments which helped improve the paper.

8 Limitations

Given the societal importance of fact-checking applications, it is important that the limitations of our method be explored. Our experimental results reveal that the surrogate entailment model often assigns truthfulness labels for entailment even when it struggles to fully grasp the relationship between the claim and the summary with evidence. This issue not only impacts the judgment of the claim label but also affects MSP during training. One potential solution is using a textual entailment model adept at managing this uncertainty or excluding such instances during training. Secondly, Llama 2's claims in the Multi-News-Fact-Checking dataset have certain flaws. Our review suggests that neutral claims might mix consistent and conflicting details. Enhancing our data creation prompts or the prompts used in the second-stage claiming could boost Llama 2's understanding.

Our model, trained on English text and topics from the Multi-News benchmarks, may not perform well in other languages without retraining. Care should be taken to ensure the model is trained on data that closely aligns with the target domain of interest, if possible, to minimize errors. Finally, our model relies on identifying relevant and trusted source documents on which to perform summarization and checking. While this document-level retrieval task is orthogonal to our research, failure to retrieve relevant documents will affect the downstream performance of the fact-checking system. If irrelevant documents are used, even true claims might be wrongly challenged. Thus, approaches should confirm that events and entities in sourced documents are directly related, employing sophisticated methods.

References

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Soc. Netw. Anal. Min.*, 13(1):30.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):821–829.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Nicolas Berlinski, Margaret Doyle, Andrew M. Guess, Gabrielle Levy, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, and Jason Reifler. 2023. The effects of unsubstantiated claims of voter fraud on confidence in elections. *Journal of Experimental Political Science*, 10(1):34–49.

D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. 2012. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag.

Andrea Ceron and Paride Carrara. 2023. Fact-checking, reputation, and political falsehoods in italy and the united states. *New Media & Society*, 25(3):540–558.

Gullal Singh Cheema, Sherzod Hakimov, Abdul Sittar, Eric Müller-Budack, Christian Otto, and Ralph Ewerth. 2022. MM-claims: A dataset for multimodal claim detection in social media. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 962–979, Seattle, United States. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, 60(2):103219.

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.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan

Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. 2022. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023a. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.

Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023b. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11).

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.

Aman Khullar and Udit Arora. 2020. MAST: Multimodal abstractive summarization with trimodal hierarchical attention. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 60–69, Online. Association for Computational Linguistics.

Sarah Kreps, R. Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1):104–117.

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

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2022. Long text and multi-table summarization: Dataset and method. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1995–2010, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mucahid Kutlu, and Yavuz Selim Kartal. 2021. Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 264–291, Cham. Springer International Publishing.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

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

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Ratish Puduppully and Mirella Lapata. 2021. Data-to-text Generation with Macro Planning. *Transactions of the Association for Computational Linguistics*, 9:510–527.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Fei Wang, Zhewei Xu, Pedro Szekely, and Muhao Chen. 2022a. Robust (controlled) table-to-text generation with structure-aware equivariance learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5037–5048, Seattle, United States. Association for Computational Linguistics.

Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. 2022b. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 939–948.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for 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*. ACM.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. 2023. Secrets of rlhf in large language models part i: Ppo.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

9 Appendix

9.1 Prompting For Multi-News-Fact-Checking Dataset

In this section, our main goal is to explain the dataset construction process. The following prompts include generating entailment, neutral, and contradiction claims from the Multi-News memorization dataset, ensuring each claim aligns with its corresponding label, and providing clear and concise claims, as depicted in Figure 6.

- **Prompt for the entailment claims**: Task: You will be provided with a summary of a news article. Your goal is to generate a list of statements derived from the summary. These statements should be definitively true based solely on the information in the summary. Example summary: The unemployment rate dropped to 8.2% last month, but the economy only added 120,000 jobs, when 203,000 new jobs had been predicted, according to today's jobs report. Reaction on the Wall Street Journal's MarketBeat Blog was swift: "Woah!!! Bad number." The unemployment rate, however, is better news; it had been expected to hold steady at 8.3%. But the AP notes that the dip is mostly due to more Americans giving up on seeking employment. You will be given a summary of a news article. Your job is to generate a list of entailment claims(true) from the summary. For example, if the summary says job growth was expected to be 100,000 jobs, but only was 80,000 jobs, one simple claim you might write could be "Job growth missed expectations." Please write a numbered list of 10 claims from this summary (numbered 1. through 10.).

- **Prompt for the neutral claims**: Task: You will be provided with a summary of a news article. Your goal is to generate a list of statements derived from the summary. These statements should not be definitively true or false based solely on the information in the summary. In other words, they should be ambiguous and require further investigation or context to determine their accuracy. Example: If the summary mentions that two celebrities are planning to get divorced, you might create a statement suggesting that their divorce might lead to significant financial and legal complications, assuming this information is not explicitly confirmed or denied in the article. Instructions: Review the provided summary. Create 10 statements based on the information in the summary. Each statement should be carefully crafted to be neither definitively true nor false based solely on the summary. Ensure that the truth or falsehood of these statements cannot be logically deduced from the summary alone. Avoid simply rephrasing or restating sentences from the summary; strive for creativity in your statement generation process. Avoid claims using statements like "may" or "could" - your claim should state things as a fact.

- **Prompt for the contradiction claims**: Task: You will be provided with a summary of a news article. Your goal is to generate a list of statements derived from the summary. These statements should be definitively false based solely on the information in the summary. Example: If the summary mentions that a black race car starts up in front of a crowd of people., you might create a statement suggesting that a man is driving down a lonely road assuming this information is explicitly denied in the article. Instructions: Review the provided summary. Create 10 statements based on the information in the summary. Each statement should be carefully crafted to be definitively false based solely on the summary. Avoid simply rephrasing or restating sentences from the summary; strive for creativity in your statement generation process. Avoid claims using statements like "may" or "could" - your claim should state things as a contradiction fact.

- **Prompt for double-check claims**: Task: You will be presented with a set of documents and one claim. Your objective is to discern the claim label based on the information in the documents. The claim labels include entailment, neutral, and contradiction. Entailment signifies that the claim is conclusively true based solely on the documents. The neutral label indicates that the claim should neither be true nor false based on the information provided. The contradiction label implies that the claim is entirely false based on the information presented in the documents.

- **Prompt for the clear and concise claims**: You will be provided a summary that a fact-

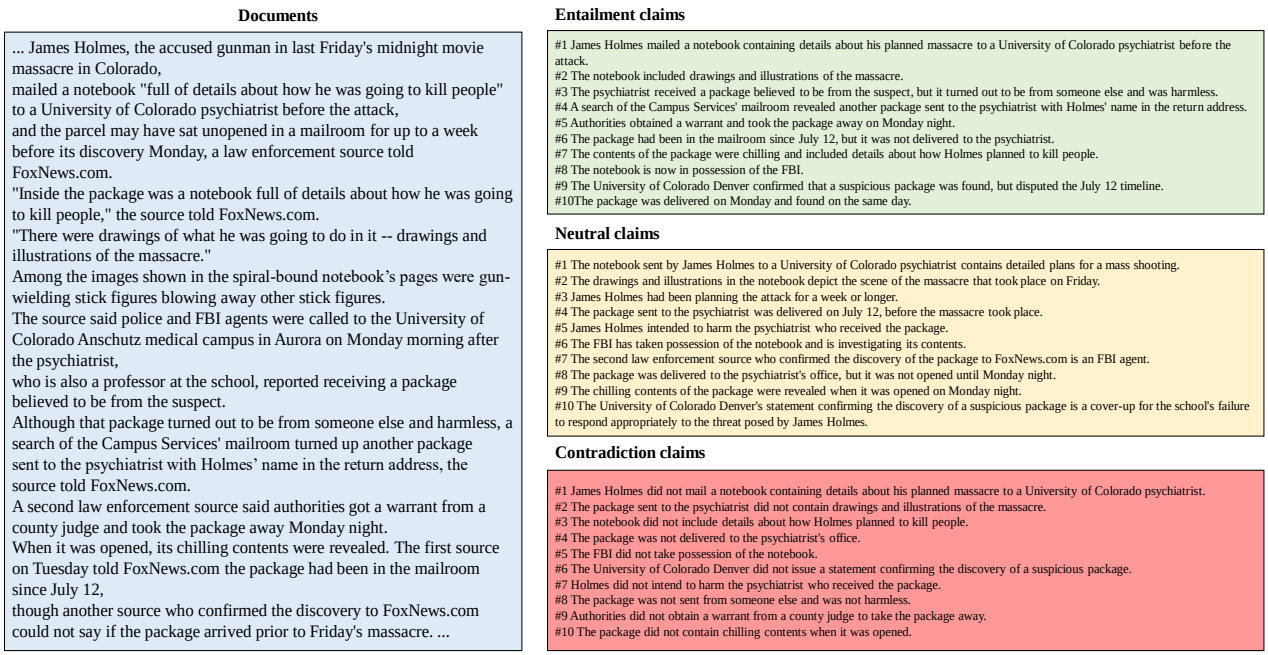

Figure 6: The prompted entailment, neutral, contradiction claims from Llama-2-70b.

checker will use for fact-checking a claim. Your task is to evaluate the provided summary using a quality assessment metric that measures whether the summary is factual and written in a clear and concise manner. Good summaries provide facts useful for fact-checking (a general claim) and are short to ease the fact-checkers job. Given a summary, your job is to provide a number from 0 to 1 that indicates your assessment of the quality of the summary. Provide the evaluation score in the format: "The quality score is <score>." The score should range from 0 to 1, where a score of 1 indicates high quality and a score of 0 signifies the lowest quality. The provided summary: <summary>"

9.2 Implementation Details

We used 4 NVIDIA A40 to run our experiments. Our model costs 180 GB and are trained for about 24 runs with a batch size of 256. In the preprocessing and evaluation parts, we use NLTK, ROUGE, and BERTScore. For claim verification, the learning rate $\in \{10^{-4}, 10^{-5}, 10^{-6}\}$ and batch size $\in \{256, 480, 512\}$.

Figure 7: The assessment of generated summaries with and without visual information.