Multimodal Multi-Document Evidence Summarization For Fact-Checking

Ting-Chih Chen

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Christopher Thomas, Chair
Lifu Huang
Ismini Lourentzou

December 1, 2023
Blacksburg, Virginia

Keywords: Knowledge Graph, Multimodal learning, Reinforcement learning and Summarization

Multimodal Multi-Document Evidence Summarization For Fact-Checking

Ting-Chih Chen

(ABSTRACT)

Fact-checking real-world claims is a time-consuming task that involves reviewing various documents to determine the truthfulness of claims. The current research challenge is the absence of a method to supply evidence that can assist human fact-checkers effectively. To solve the research challenge, we propose the **MetaSumPerceiver** model designed to create claim-specific summaries for fact-checking. The **MetaSumPerceiver** model, a dynamic perceiver-based model, takes inputs in the form of documents, images, and a claim, with the objective of assisting in fact-checking tasks and handling inputs of varying lengths from multiple modalities. To train this model, we use a novel reinforcement learning-based entailment objective to generate summaries that offer evidence distinguishing between different truthfulness labels.

To assess our model's effectiveness, we introduce the **KG2Claim** approach to generate multimodal multi-document claims. This approach integrates information from multimodal multi-document sources into the knowledge graphs. Our main objective is to examine whether the multimodal multi-document claims align with the information in articles. The findings from **MetaSumPerceiver** show that more than 70% of our claims are entailment claims. This validates that **KG2Claim** effectively generates claims that entail the information from multimodal multi-document sources. Subsequently, we conduct evidence summarization experiments on an existing benchmark and a new dataset of multi-document claims that

we contributed. Our approach surpasses the state-of-the-art method by 4.2% in the claim verification task on the MOCHEG dataset and demonstrates strong performance on our Multi-News-Fact-Checking dataset.

Multimodal Multi-Document Evidence Summarization For Fact-Checking

Ting-Chih Chen

(GENERAL AUDIENCE ABSTRACT)

Social media constantly provides us with a mix of images, videos, and text from various sources. Human fact-checkers, in their efforts to verify the accuracy of this information, often spend 2-3 hours on a single post, carefully examining all the content. Human fact-checkers not only verify the information's relevance to its sources but also need to assess its accuracy in terms of entailment, neutrality, and contradiction. These three categories ensure that the statement is appropriately related to the sources. To assist human fact-checkers, we propose an evidence summarization model, **MetaSumPerceiver**, designed to create concise summaries tailored to specific claims. **MetaSumPerceiver**, accepting inputs in the form of documents, images, and a claim, aims to facilitate fact-checking tasks.

To evaluate the effectiveness of the **MetaSumPerceiver** model, we also introduce the **KG2Claim** approach. This approach employs knowledge graphs and multimodal coreference resolution, efficiently integrating information from multimodal multi-document sources. The results indicate that more than 70% of our generated claims are entailment claims, signifying that the majority of claims are related to multimodal multi-document sources. Subsequent experiments of **MetaSumPerceiver** were conducted on both an existing benchmark and a new dataset of multi-document claims, which we contributed. The results indicate a notable improvement over the state-of-the-art approach, achieving a 4.2% better performance in the claim verification task on the MOCHEG dataset. Moreover, our approach

demonstrated robust performance on the Multi-News-Fact-Checking dataset. This thesis contributes an evidence summarization model aimed at aiding human fact-checkers in assessing the truthfulness of claims through concise summaries tailored to specific claims. Furthermore, we introduce a practical multimodal multi-document claim generation approach that consolidates knowledge from different documents.

Dedication

To my parents, family, advisor, and lab members, I am truly grateful for the continuous support and faith you have shown me throughout my thesis journey. Your unwavering encouragement, direction, and teamwork have been the primary motivators for my achievements.

Acknowledgments

I extend my heartfelt gratitude to my advisor, Chris Thomas, for his insightful guidance and unwavering support, which played a pivotal role in shaping the direction of this thesis. I am also thankful to my family for their constant encouragement and belief in my abilities, which provided the foundation for my academic journey.

Contents

List of Figures

List of Tables

List of Abbreviations

LLMs Large Language Models

NLP Natural Language Processing

PPO Proximal Policy Optimization

NLP is a branch of machine learning that empowers computers to understand, manipulate, and interpret human language.

LLMs are advanced deep learning algorithms capable of executing a diverse range of tasks related to natural language processing (NLP).

PPO algorithm belongs to the category of policy gradient methods. PPO algorithms offer advantages similar to trust region policy optimization (TRPO) but stand out for being simpler to implement, more versatile, and exhibiting improved sample complexity.

Chapter 1

Introduction

1.1 Research Background

Presently, a significant portion of news circulating on social media platforms like Facebook, Reddit, and Twitter is characterized by disinformation and misinformation. This prevalence of false information can significantly misguide readers, leading them towards extreme viewpoints and fostering distrust in government actions. Notably, this phenomenon is particularly pronounced in the lead-up to elections, as witnessed during the 2016 US presidential election, where fake news proliferated across social media platforms, impacting public sentiment. Similarly, the Brexit referendum saw both "Leave" and "Remain" campaigns accused of spreading disinformation, particularly regarding economic consequences and immigration. Amidst this backdrop, reliable news sources, such as National Public Radio and The Wall Street Journal, are scarce, emphasizing the need for a robust fact-checking method. This method should not only accumulate knowledge from diverse articles covering the same story but also possess the capability to comprehend information from different modalities. Such a comprehensive approach is essential to effectively combat the spread of disinformation and misinformation in the age of social media.

Fact-checking claims on social media platforms poses a significant challenge due to the large volume of news claims constantly being posted without sufficient methods for verification [1]. Research [2] indicates that manually verifying all aspects of a 200-word claim can require

up to 4 hours of dedicated effort because human fact-checkers must find supporting evidence which could require reviewing multiple sources, including articles, images, videos, etc. Further, despite the exceptional capabilities of large language models (LLMs) [3] in natural language processing (NLP) tasks, they still struggle to understand events across documents. The limitation [4, 5, 6, 7, 8] is particularly problematic in the context of fact-checking, where a comprehensive understanding of events is crucial for accurate assessment. Recognizing the potential for LLMs to inadvertently generate misleading information, we underscore the importance of developing models specifically tailored to handle the complexity of cross-document event comprehension. As disinformation [9, 10] continues to proliferate, addressing this gap in LLM capabilities becomes imperative to ensure reliable fact-checking and information verification. To solve this issue in the fact-checking task, our method mainly focuses on generating claim-specific summaries to assist this task following the evidence retrieval, claim verification, and justification production.

1.2 Limitations and Challenges

There is a pressing requirement for tools that can efficiently combine multimodal multi-document information and provide concise evidence summaries for fact-checkers. Currently, there is limited research exploring the integration of multimodal information, especially in the fact-checking task [11]. Existing research [12, 13, 14] relying on summarization for fact-checking is ineffective because it fails to extract evidence from the sources. These approaches lack the ability to comprehend entire articles and produce specific evidence to substantiate their claims. Moreover, prior research [15] has shown existing systems are unable to effectively handle multimodal data.

Multimedia event extraction [16, 17, 18, 19, 20] offers a potential solution to integrate infor-

mation from multimodal sources. It focuses on extracting events and their associated details from various modalities simultaneously. The primary goals include classifying events into pre-defined event types and identifying arguments for each event, grounded in text entities or image objects. This task is demanding due to the complementary nature of information from different modalities. While multimedia event extraction can identify events and entities across various modalities, it falls short as an effective solution for the fact-checking task. The rationale behind this is that humans find multimodal information to be less efficient as a representation for quickly comprehending content.

Multimodal summarization [21, 22, 23, 24] offers a promising solution to the challenge of condensing evidence. By integrating information from various sources, such as text, images, videos, and audio, it enables the generation of summaries that enhance people's understanding of diverse content. This is a challenging task because each modality might contribute complementary information, e.g. bar chart image with other relevant facts mentioned in the text. Current methods [25, 26, 27] typically generate intuitive summaries using multimodal information. However, our aim differs. We do not use intuitive summaries for fact-checking since they lack the specific details needed to verify events or entities. Our goal is to efficiently distill claim-specific evidence useful for fact-checking across various modalities.

According to the current limitations mentioned above, we are still missing the method of generating evidence for the fact-checking task. We employ multimodal summarization techniques to create a model that generates claim-specific evidence for the fact-checking task. Furthermore, we follow multimedia event extraction lead in constructing a method that harnesses the power of these approaches, aiming to create a comprehensive framework for handling diverse sources of information. We envision an approach that seamlessly integrates textual, visual, and potentially auditory information to furnish the claim-specific summary for the fact-checking task. This integration will enable a more robust and contextually rich

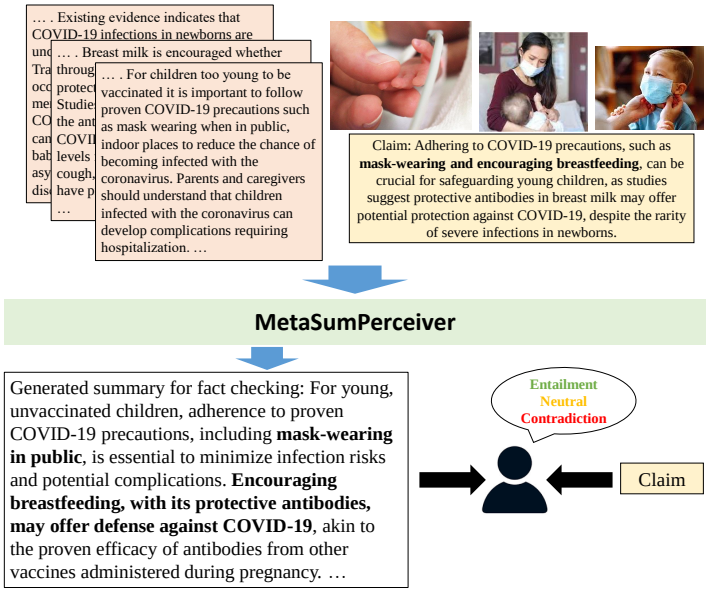

Figure 1.1: Overview of MetaSumPerceiver: Using inputs such as documents, images, and claims, MetaSumPerceiver generates summaries to facilitate fact-checking. In this example, the summary for fact-checking provides evidence and establishes that the claim in question is entailed by the evidence.

understanding of the content, breaking down the barriers between different modalities.

1.3 Contributions

To overcome the challenge of fact-checking with multimodal multi-document sources, we propose the **MetaSumPerceiver** model in Figure 1.1, 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 [28, 29] with a summarization model [30]. 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. Then, to train **MetaSumPerceiver** to generate summaries useful for human fact-checking, we assess the utility of our

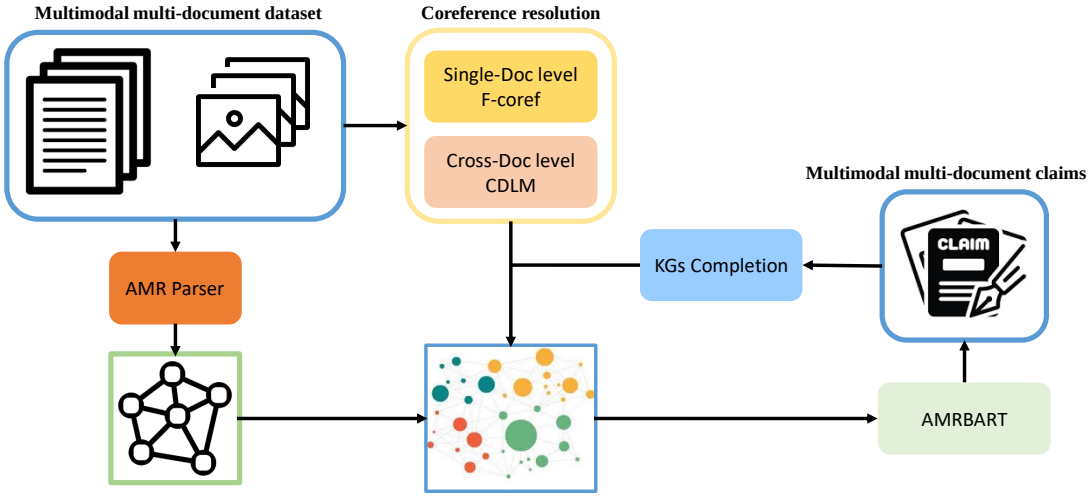

Figure 1.2: Pipeline of KG2Claim: The elements encompass an AMR parser, single-document and cross-document coreference resolution, Knowledge graphs completion with LLMs, and AMRBART, generating the multimodal multi-document claims.

summaries at performing entailment [31, 32, 33], 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. In order to support research on the task of multimodal multi-document fact-checking, we also introduce the **KG2Claim** approach outlined in Figure 1.2. This approach takes a multimodal multi-document knowledge graph as input and aims to create claims that incorporate this diverse information through multimodal coreference resolution [34, 35, 36].

To assess the efficiency of the **MetaSumPerceiver**, we employ our **KG2Claim** method for the fact-checking task, evaluate it against the MOCHEG benchmark [37], and introduce a new benchmark called Multi-News-Fact-Checking. Multi-News-Fact-Checking benchmark involves claims and entailment labels supported by evidence from multiple documents. Our

results indicate significant enhancements compared to existing baselines.

The major contributions of this thesis are as follows:

- We present an innovative approach for multimodal multi-document summarization specifically designed for fact-checking applications.
- We introduce a claim generation method tailored for disinformation detection tasks, with a focus on handling multimodal multi-document information.
- We release the Multi-News-Fact-Checking dataset, to support the multi-document fact-checking summarization task.
- We perform detailed experiments and ablations of our approach and loss functions which clearly demonstrate the superiority of our approach over existing methods.

Chapter 2

Review of Literature

2.1 Knowledge Representation

Abstract Meaning Representation (AMR) [38, 39, 40] functions as a robust semantic representation language, employing rooted, labeled, directed, and acyclic graphs to encapsulate entire sentences. This representation brings forth two pivotal advantages. Firstly, AMR serves as a conduit for semantic representation, adept at capturing the intricate structural nuances within sentences, thereby delineating the relationships that underpin various entities. Secondly, AMR seamlessly integrates semantic role labeling, shedding light on the nuanced roles assumed by different words in a sentence, such as agents, patients, or locations.

The strength of AMR lies in its holistic approach to language representation. AMR not only enhances semantic clarity but also facilitates a more profound understanding of the relationships inherent in linguistic expressions. This depth is particularly valuable when unraveling complex narratives or dissecting intricate layers of meaning within the text. Comparatively, when juxtaposed with Information Extraction (IE) [41, 42], AMR emerges as a comprehensive information tool. Unlike IE, which is confined to extracting predefined information from textual sources and struggles with untrained sentences, AMR exhibits the versatility to analyze diverse sentences through its nuanced semantic representation.

Harnessing the advantages of AMR, this thesis employs IBM AMR parser [1] for extracting textual sources. By doing so, it not only dissects the relationships between events and entities but also benefits from the inherent depth that AMR brings to the semantic analysis of linguistic expressions.

2.2 Coreference Resolution

Coreference resolution is the task of finding all expressions that refer to the same event and entity in a text. It is an important step for a lot of higher-level NLP tasks that involve natural language understanding such as document summarization, question answering, and information extraction. Recent research has introduced innovative approaches to sentence-level coreference resolution [43, 44] and document-level coreference resolution [45, 46]. The primary technology entails detecting candidate mentions, encoding them into vector representations, and identifying coreference relations by employing a Multilayer Perceptron (MLP) classifier to process the representations of each mention and past entities [47]. Notably, Joshi et al. [48] established the SpanBERT model, achieving state-of-the-art performance in coreference resolution. Furthermore, Yao et al. [49] developed a model inspired by SpanBERT, designed to learn and integrate multiple representations from both event alone and event pair.

We implement coreference resolution at both the single-document and cross-document levels. For single-document coreference resolution, we employ F-coref [34] to identify the same entities on a sentence-by-sentence basis. Subsequently, for the cross-document coreference resolution, we utilize CDLM [36] to extract coreferent entities. This crucial step serves to establish connections between identical entities within the knowledge graphs. Moreover, it

[1] https://github.com/IBM/transition-amr-parser

enables tracking the specific document in which these entities are mentioned.

2.3 Knowledge Graph Completion

In knowledge graphs (KGs), latent relationships between events and entities often result in unknown connections, necessitating approaches to comprehend the missing information; this gap is addressed through knowledge graph completion [50, 51, 52], a method that predicts and fills in missing relationships or edges between entities, thereby enhancing the overall comprehensiveness of the knowledge graph.

Bordes et al. [53] invented TransE to construct entity and relation embeddings by treating relations as translations from head entity to tail entity. Inspired by Mikolov et al. [54], TransE learns vector embeddings for entities and relationships, placing them in $\mathbb{R}^k$. The fundamental concept behind TransE is that the relationship between two entities is akin to a translation between their embeddings, expressed as $h + r \approx t$ when (h, r, t) holds, where h represents the head entity, t represents the tail entity and r represents the relationship from the head entity to tail entity, respectively. However, due to challenges in modeling 1-to-N, N-to-1, and N-to-N relations, Wang et al. [55] introduced TransH to allow entities to have distinct representations in different relations. Incorporating BERT [56] into the knowledge graph completion task, Yao et al. [57] developed KG-BERT. This model takes the entity and relation descriptions of a triple as input, computes the scoring function for the triple using the KG-BERT language model, and predicts unknown relationships. Zhang et al. [58] incorporate the helpful KGs structural information into the LLMs, aiming to achieve structural-aware reasoning in the LLMs. They first transfer the existing LLMs paradigms to structural-aware settings and propose a knowledge prefix adapter to predict the unknown relationships.

To address this challenge in KGs, we utilize Vicuna [59] to identify hidden relationships within each claim. Furthermore, we incorporate additional claims to elucidate latent connections in the KGs, enhancing the information within knowledge graphs. This process aids in a more comprehensive understanding of events and entities during the generation of multimodal multi-document claims.

2.4 Perceiver

The perceiver architecture [60] in Figure 2.1 enables scaling transformers to input sequences of arbitrary lengths, by reducing the memory footprint in standard self-attention. Perceiver is an architecture grounded in attentional principles, designed to handle high-dimensional inputs and multimodal combinations without relying on domain-specific assumptions. It employs a cross-attention module to map a high-dimensional input byte array to a fixed-dimensional latent bottleneck. Subsequently, it processes this bottleneck through a deep stack of Transformer-style self-attention blocks in the latent space. The perceiver engages in an iterative process of attending to the input byte array by alternating between cross-attention and latent self-attention blocks.

Follow-up works, such as Perceiver IO [28], 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 the GLUE language benchmark [61], Sintel optical flow estimation [62], and others all without the need for explicit multiscale correspondence mechanisms.

Uni-Perceiver v2 [63] stands out as the first generalist model capable of efficiently handling major large-scale vision and vision-language tasks. Notably, Li et al. [63] can directly manage

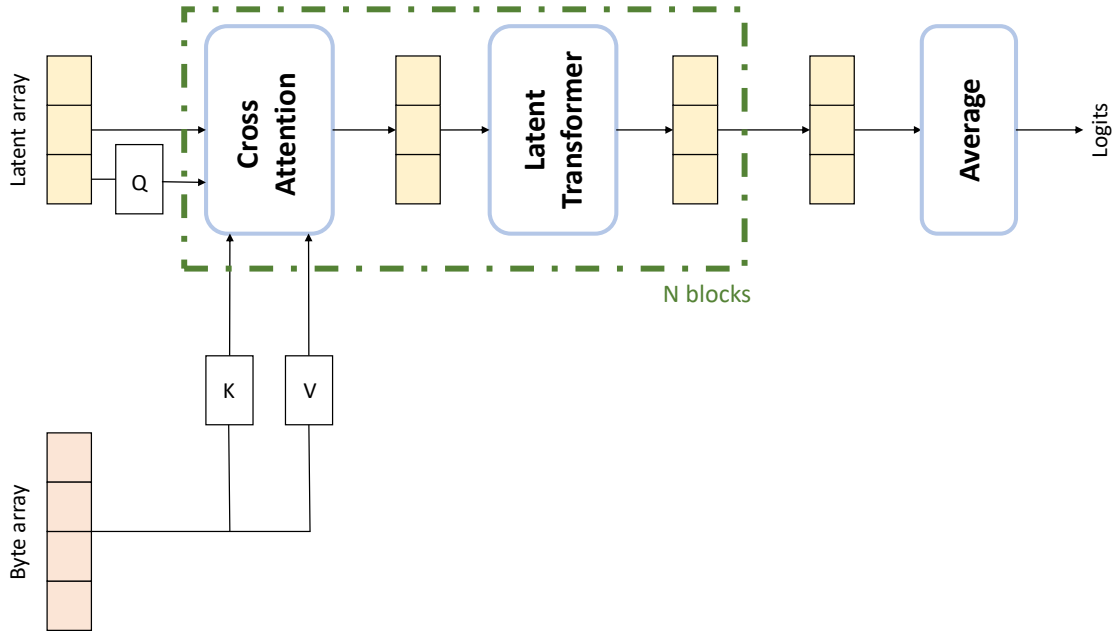

Figure 2.1: The perceiver architecture.

downstream tasks without requiring task-specific adaptation, such as image classification, object detection, image-text retrieval, and image captioning. Our method relies on Li et al. [63] 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.

2.5 Multimodal Summarization

Recently, a number of approaches have been proposed for generating summaries of multimodal content. Rott and Červa [23] use input audio to generate textual summaries. Sah et al. [24] extract textual summaries from annotated and summarized videos. Junnan et al. [64], Zhu et al. [65] propose an integrated approach which utilizes both textual and visual modalities as inputs and produce multimodal outputs summarizing text and video. Palaskar

et al. [66] propose to generate abstractive summaries from open-domain videos. Despite this recent progress, existing models continue to struggle with capturing complementary information from multiple modalities. Unlike prior work in this space, we seek to generate textual summaries of evidence from multiple modalities for purposes of fact-checking.

2.6 Learning from Feedback

Recent advancements in LLMs have revolutionized the AI landscape [3, 67, 68, 69]. However, because they are mostly trained on data scraped from the web LLMs sometimes produce undesired outcomes, including generating biased or harmful content [70]. Recognizing the importance of aligning LLMs with human values, has led to efforts in supervised fine-tuning (SFT) with ethical guidelines [71]. While these efforts demonstrate the potential of integrating human feedback into training using reinforcement learning for user-tailored tasks [72, 73], 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) [74, 75].

2.7 Training Strategy: from BERT to DeBERTa V3

Devlin et al. [56] employs two key unsupervised training tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP), which collectively contribute to the model's ability to understand contextualized word representations and relationships between sen-

tences. The MLM task involves randomly masking some words in a sentence and training the model to predict the masked words based on the context of the surrounding words. This helps BERT learn bidirectional contextual representations, capturing the meaning of words in the context of the entire sentence. The NSP task, on the other hand, involves predicting whether one sentence follows another in a document. This task encourages the model to understand the relationships and coherence between sentences, enabling BERT to grasp the broader context of a document. By combining these tasks during pre-training, BERT learns rich contextualized representations that make it highly effective for a wide range of natural language processing tasks.

To enhance performance, Clark et al. [76] suggests replacing the pretraining task with the "Replaced Token Detection" (RTD) task, akin to generative adversarial networks (GANs). In the RTD task, the objective is to discern whether a given token is original or not. Results indicate that weight sharing outperforms the MLM task because the RTD task updates only the input token embedding in the discriminator. DeBERTa V3 in Figure 2.2 integrates the DeBERTa model with the RTD task, enhancing relative position embedding and emphasizing absolute position embedding in the final layer. Comparative results demonstrate DeBERTa's superiority over BERT. Moreover, DeBERTa V3 retains the DeBERTa model with the RTD task, achieving the best performance in GLUE with an impressive 91.37% average score according to experiments. In our work, we primarily employ DeBERTa V3 as our entailment model to help determine whether a claim is entailed by a summary or not.

2.8 Prompting from LLMs

The direct impact of a prompt or prompt strategy on model outputs, as well as the modification of LLMs' billions of parameters during re-training, are both active areas of re-

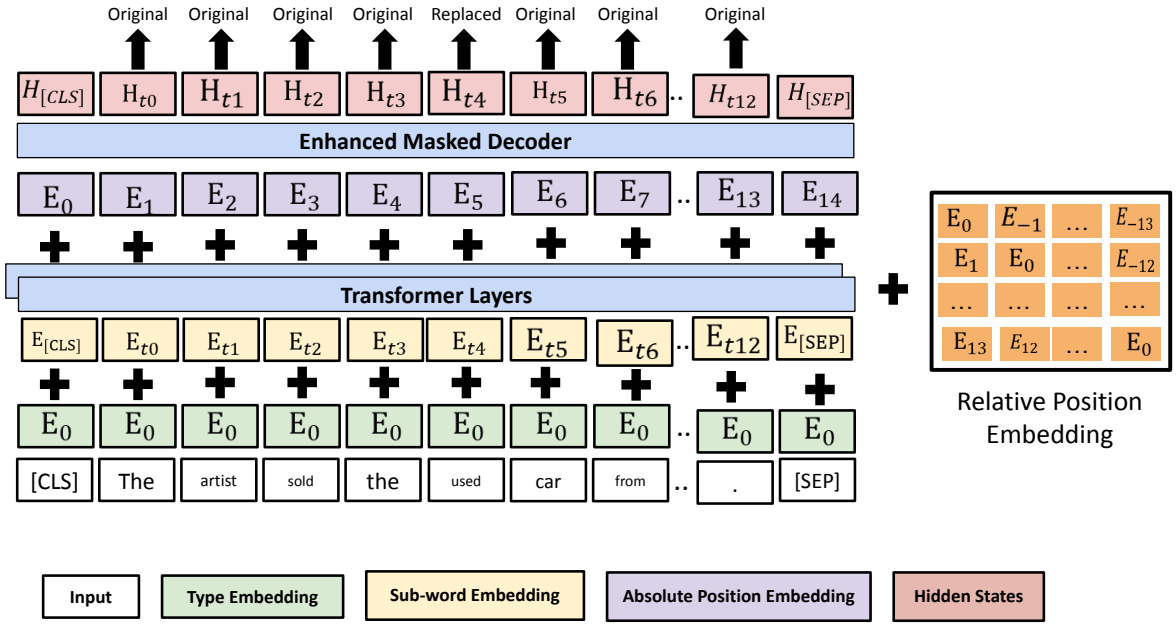

Figure 2.2: Overview of training strategy in DeBERTa V3: This figure involves passing sub-word embedding which is token embedding through several transformer layers, adding it to position embedding. It then goes through another transformer layer, followed by training using replacing token detection. The position embedding is divided into absolute and relative components. Absolute follows the original BERT, emphasizing absolute word positions. Relative, on the other hand, emphasizes relative positions. This approach highlights absolute word positions in the final layer.

search [77, 78]. Recent research [79, 80, 81] provides some insights into effective prompt design strategies. Brown et al. [82] demonstrated that examples significantly enhanced GPT-3's performance across tasks such as question answering and language translation. We focus on designing effective prompts for fact-checking. Additionally, we release our prompted results as a pseudo-labeled dataset to establish a benchmark on this task.

Chapter 3

Methodologies

3.1 Evidence Summarization Model (MetaSumPerceiver)

3.1.1 Model Training Strategy

We explain the details of our approach, **MetaSumPerceiver** in Figure 3.1, as well as how we construct our Multi-News-Fact-Checking dataset as illustrated. 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 **MetaSumPerceiver**. 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.

For the textual data, we use BART [30] [1] to obtain text embeddings following [83, 84]. 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. We use CLIP (ViT-G-14) [85] 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.

Our goal is to generate a textual summary of a set of multimodal documents that enables a

[1] https://huggingface.co/facebook/bart-large-cnn.

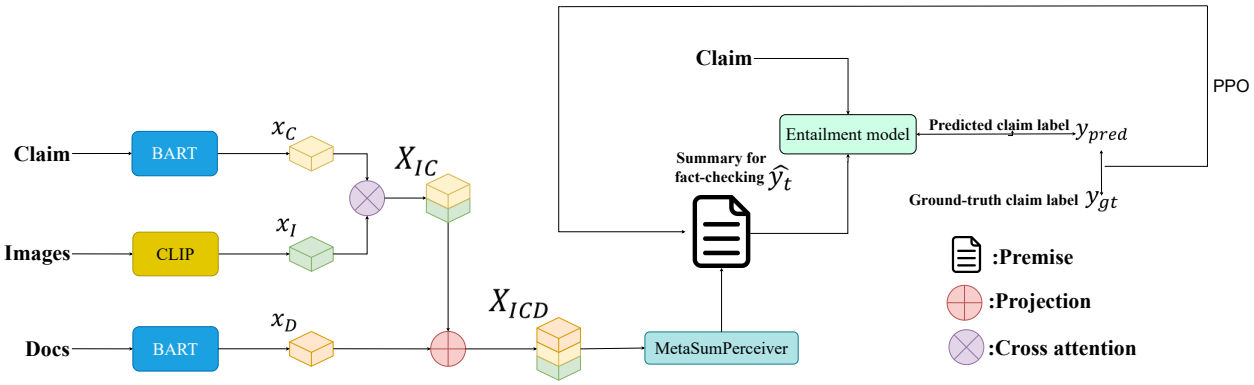

Figure 3.1: Overview of MetaSumPerceiver: This figure illustrates the process of generating a summary for fact-checking using MetaSumPerceiver, integrating a fixed entailment model for accurate truthfulness labeling. Furthermore, it highlights how PPO is employed to continually refine the summary during the fact-checking process.

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:

$$X_{IC} = ATTN(Q_{x_C}, K_{x_I}, V_{x_I}) \,, \tag{3.1}$$

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 **MetaSumPerceiver**.

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:

$$X_{ICD} = \left[proj(X_{IC}, \theta)^\intercal, X_D^\intercal \right]^\intercal \,, \tag{3.2}$$

where X_{ICD} will be the input to **MetaSumPerceiver**. Prior to training our full model, we pretrain our attention block and summarization model using the Multi-News dataset's

human written summaries using the cross-entropy loss function:

$$\mathcal{L}_{\text{sum}} = -\sum_{t=1}^{T}\sum_{i=1}^{N} y_{t_i}\log(\hat{y}_{t_i}), \tag{3.3}$$

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.

3.1.2 Reward Model for Fine-tuning the Summarizer

To enhance the summarizer's ability to produce summaries that provide the evidence needed for 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 (RL) 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 3.2. 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 [86], Fever-NLI [87], and Adversarial-NLI (ANLI) [88], encompassing a total of 763,193 premise-claim pairs. Leveraging this dataset, we fine-tuned DeBERTa V3 [89] for the task of entailment classification using cross-entropy loss. Serving as an entailment classifier, this model achieves accuracy rates

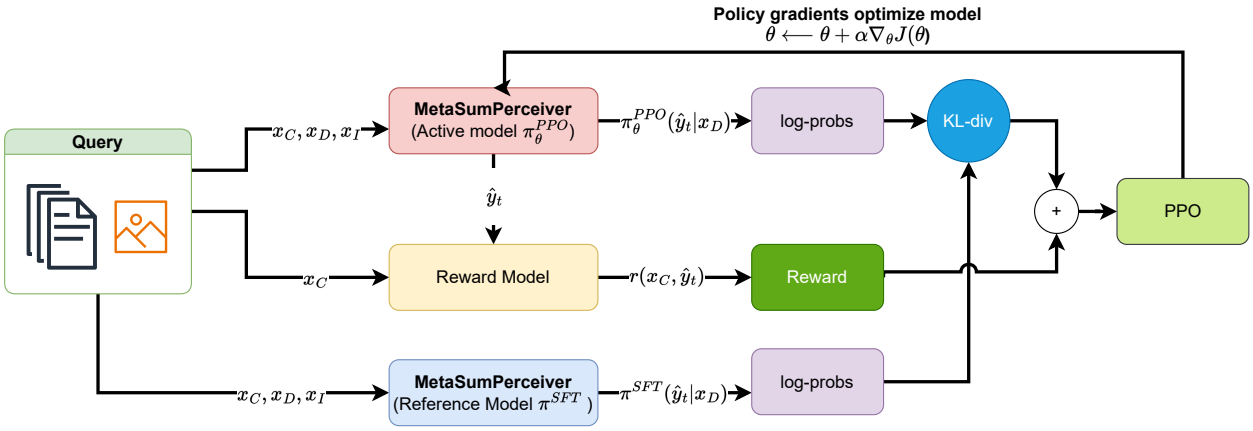

Figure 3.2: The Proximal Policy Optimization (PPO) workflow begins with the summarizer creating a response based on the input query. The reward model then evaluates this query-response pair that outputs a scalar reward. Meanwhile, the process calculates the KL-divergence based on the likelihood of token sequences in the response using both an active model being fine-tuned currently and a pre-trained reference model. The KL-divergence serves as a reward measure, ensuring responses from the active model are aligned with the reference model. Conclusively, PPO updates the active model's parameters, relying on the result of the reward model's output and the KL-divergence's value.

of 90.3%, 77.7%, and 57.9% in the MultiNLI, Fever-NLI, and ANLI evaluation datasets, respectively. 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:

$$r(x_C, \hat{y}_t) = P(y_{gt}|x_C, \hat{y}_t) - 0.5 * \Sigma_{y_{gt} \neq y_{pred}} P(y_{pred}|x_C, \hat{y}_t), \tag{3.4}$$

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.

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}. This cumulative reward is described as follows:

$$r_{total} = r(x_C, \hat{y}_t) - \eta KL(\pi_{\phi}^{PPO}(\hat{y}_t|x_D), \pi^{SFT}(\hat{y}_t|x_D)), \tag{3.5}$$

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. **MetaSumPerceiver** 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:

$$\theta \longleftarrow \theta + \alpha \nabla_{\theta} J(\theta), \tag{3.6}$$

where α and J_{θ} denote the learning rate and the expected return under policy π_{θ} from the model, respectively.

3.1.3 Multi-News-Fact-Checking Dataset

In order 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 [90], which contains sets of multiple text documents

along with human-written summaries of each set. Because the Multi-News dataset doesn't have claims specifically made for fact-checking and lacks images for the news articles, we use Llama-2-70b [3]. We ask it to create labeled claims from each set of documents and get news images from Google. In each group of Multi-News documents, we use the human-written multi-document summary to generate 30 claims (ten of each type), giving us a dataset of 1,291,168 labeled claims. Additionally, we collect 111,905 images for our multimodal multi-document dataset. The prompts include sections with a task description, example, and instructions, which are fully detailed in appendices A.1.

3.2 Multimodal Multi-document Claims Generation Method (KG2Claim)

3.2.1 Multimodal Coreference Resolution

We have a two-step approach for the coreference resolution task. The initial step involves single-document coreference resolution, where we utilize F-coref [34]. Subsequently, the second step focuses on cross-document coreference resolution, where we employ CDLM [36]. In the context of single-document coreference resolution, F-coref emerges as a Python package designed for swift, precise, and user-friendly English coreference resolution. Drawing inspiration from s2e [91], F-coref introduces parallelism through batching, thereby reducing unnecessary computations like padded tokens. Similar to other neural coreference models, F-coref evaluates each pair of spans in the text for potential co-reference.

The architecture of F-coref encompasses three key components: (1) Longformer [92], serving as a contextualized encoder; (2) a parameterized mention scoring function denoted as f_m; and (3) a parameterized pairwise antecedent scoring function labeled as f_a. To determine

the coreference likelihood between any pair of spans, the model initiates by encoding the text through Longformer, resulting in vectors $x_1, ..., x_n$. Subsequently, for each potential span $q = (x_k, x_l)$, the mention scoring function $f_m(q)$ evaluates the probability of q (the "query") being a mention. For a given pair of spans $c = (x_i, x_j)$ and $q = (x_k, x_l)$, where c ("candidate") precedes q, the pairwise antecedent scoring function $f_a(c, q)$ assesses the likelihood of c being an antecedent of q. In practical terms, to mitigate the complexity of $O(n^4)$, the antecedent function scores only the top λT spans with the highest mention scores (where T represents the number of tokens). Ultimately, the final pairwise score for a coreference link between c and q comprises the scores indicating the likelihood of q and c being mentioned, along with the probability of c being an antecedent of q:

$$F(c, q) = \begin{cases} f_m(c) + f_m(q) + f_a(c, q) & c \neq \varepsilon \\ 0 & c = \varepsilon, \end{cases} \tag{3.7}$$

where ε is the null antecedent. The computation of f_m and f_q for the entire sequence can be efficiently batched.

In the cross-document coreference resolution task, we employ CDLM [36] in Figure 3.3, a pretraining approach designed for multi-document language modeling. This method incorporates two key concepts: (1) pretraining over sets of related documents with overlapping information and (2) pretraining utilizing a dynamic global attention pattern over masked tokens to reference the entire cross-text context. During pretraining over related documents, CDLM focuses on training the model on sets of documents that revolve around the same topic. This strategy aims to enhance the model's ability to understand cross-text mapping and alignment, contributing to improved unmasking.

To facilitate effective contextualization across multiple documents, CDLM leverages trans-

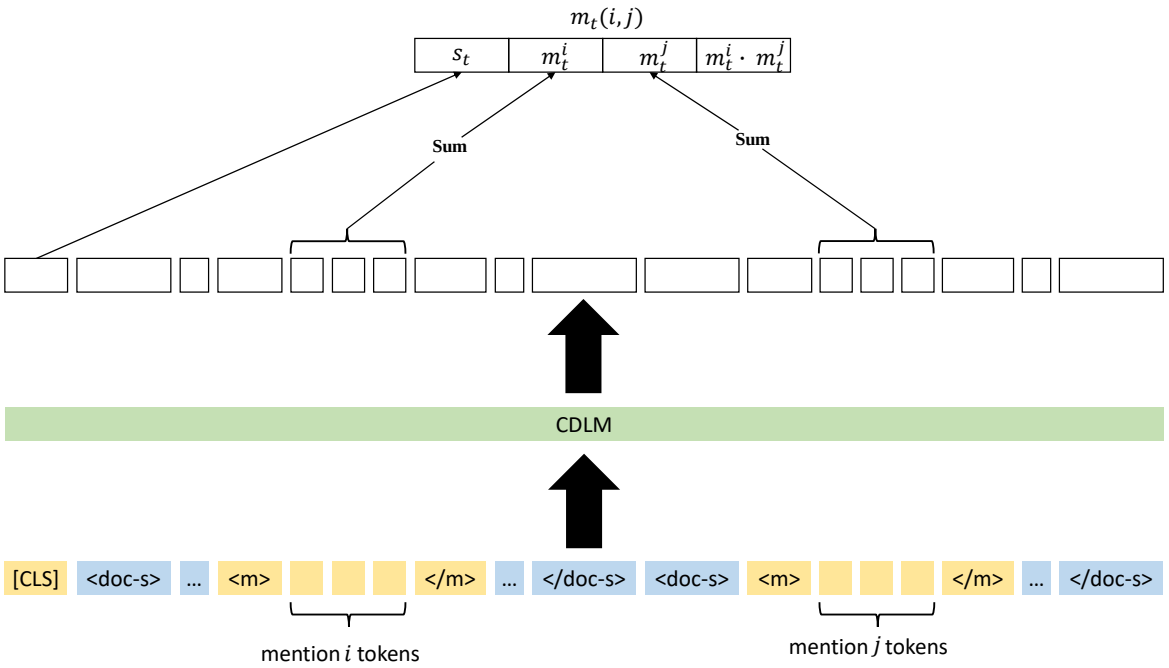

Figure 3.3: The CDLM model utilizes pairwise mention representation for coreference resolution. m_t^i, m_t^j and s_t are the cross-document contextualized representation vectors for mentions i and j, and of the [CLS] pairwise-mention representation. The tokens colored in yellow represent global attention, and the tokens colored in blue represent local attention.

former models with linear scalability concerning input length, building upon the longformer model. The model processes input by concatenating related documents using dedicated document separator tokens, <doc-s> and </doc-s>, to demarcate document boundaries. Employing a masking strategy akin to BERT [56], approximately 15% of tokens in each training example are randomly chosen to be masked. However, CDLM's pretraining approach aims to predict each masked token by considering the entire document set and assigning them global attention weights. This unique approach enables the longformer to contextualize information across documents and handle long-range dependencies within documents. Ultimately, CDLM is employed to accomplish cross-document coreference resolution.

Upon completing coreference resolution in both single-document and cross-document scenarios, we integrate coreference relationships (:coref) into the KGs. These links signify the

coreference relationships between the events and entities, aiding in the generation of claims that incorporate information from multiple documents. Figure 3.4 illustrates the multimodal multi-document knowledge graph. Colors signify distinct documents, and the connection between nodes of different colors indicates coreference for a shared event or entity.

3.2.2 Knowledge Graph Completion with LLMs

To uncover latent relationships within multiple documents, we utilize Vicuna [59] for knowledge graph completion. Vicuna is a cutting-edge model designed for natural language understanding tasks. Leveraging advanced techniques, it excels in tasks such as text summarization, sentiment analysis, and language translation. Its robust architecture ensures high-performance outcomes across diverse applications.

The process involves inputting multimodal multi-document claims generated by **KG2Claim** into Vicuna. Subsequently, prompts are employed to extract latent information from these claims. This ensures the incorporation of latent information derived from such claims. Furthermore, the completion claims are inserted into knowledge graphs to enhance available resources.

3.2.3 Knowledge-Driven Claim Generation

To produce multimodal multi-document claims from Knowledge Graphs (KGs), we devise a traversal method specifically tailored for KGs featuring coreference relationships (:coref). The traversal method is detailed in Algorithm 1. Additionally, we utilize the AMRBART model [93] to generate multimodal multi-document claims based on the information extracted from the knowledge graphs. AMRBART is a BART [30] model pretrained on AMR graphs by introducing graph self-supervised training. AMRBART utilizes two graph auto-

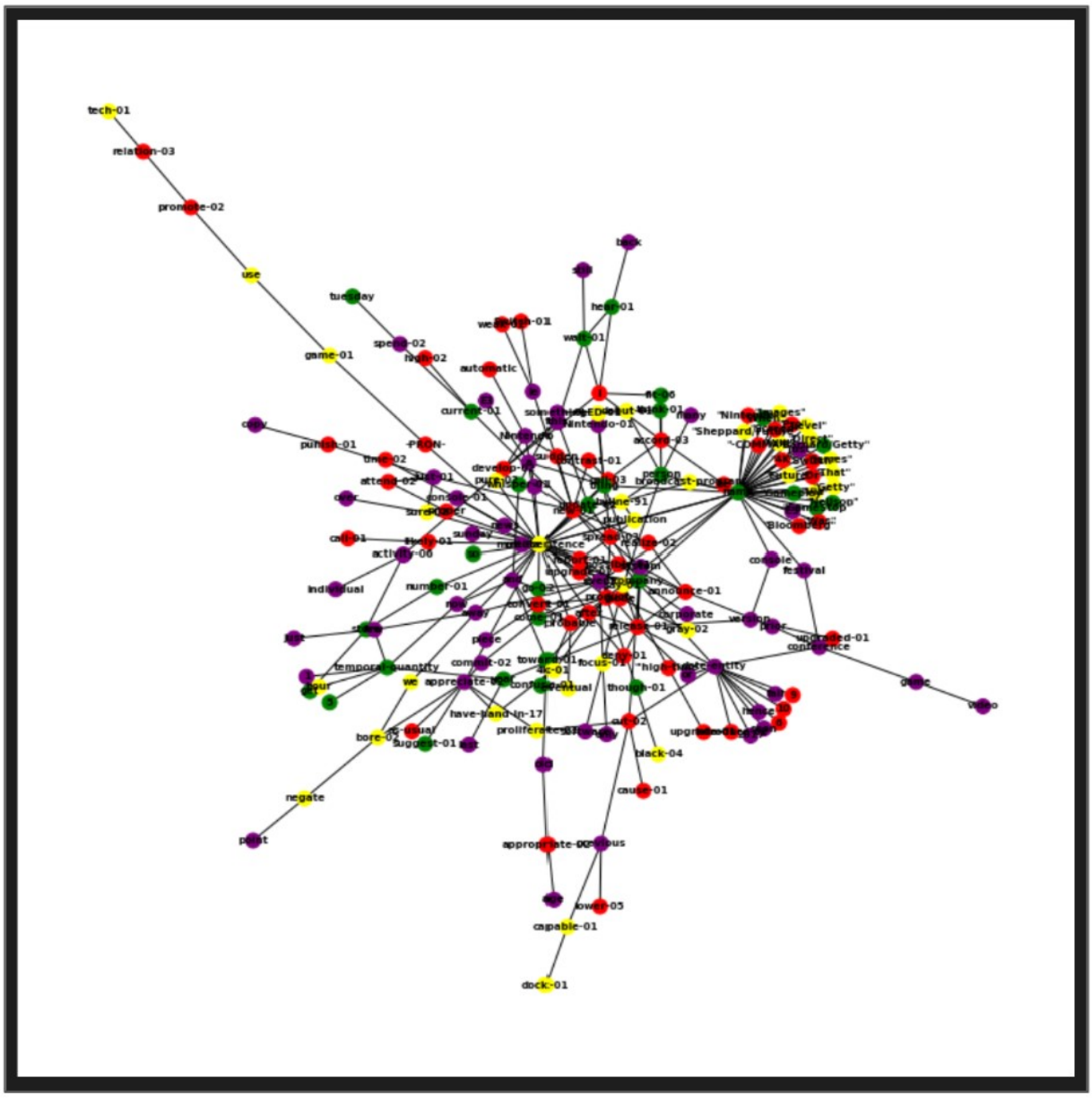

Figure 3.4: The knowledge graph is depicted with color-coded elements denoting multi-document sources. In our research, a maximum of four documents is utilized.

encoding strategies for graph-to-graph pre-training and integrating text and graph information through four tasks.

We feed the traversal results into AMRBART to generate the text. The **KG2Claim** method is a text generation method for our detecting disinformation project. We totally generated 816,115 multimodal multi-document claims in the NewsStories dataset. This method focuses on detailing the processing of text representation and traversal algorithms involving coreference relationships.

Algorithm 1 Traversal algorithm

```
Input: Graph G includes sets of (h,r,t) triple
Outputs: Set of triples with the coreference relationships (:coref)
Search all predicate nodes P that have :ARG relationships
for p ∈ P do
    Traversal results = CorefDFS(G, p)
function CorefDFS(Graph, start, visited=None)
    visited = set()
    Stack=[start]
    while Stack do
        vertex = Stack.pop()
        if vertex not in visited then
            visited.append(vertex)
            neighbors = sort(Graph[vertex], key=:coref, reverse=True)
            Stack.append(neighbor for neighbor in neighbors if neighbor not in visited)
    return Stack
```

Chapter 4

Results

4.1 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 multimodal multi-document fact-checking.

Specifically, we employed three *fixed* entailment models, namely DeBERTa V3 [89], Llama-2-70b [3], and stance detection model [37], as our surrogate "human" fact-checkers. The goal of these models is to predict the entailment label of a claim given our generated summary. Importantly, we do not finetune these models with our generated summaries so as to not bias the models towards linguistic or stylistic patterns of the summarizer. As depicted in Tables 4.1 and 4.2, our method exhibits superior performance, achieving a SOTA 48.2 F-score in the MOCHEG dataset. Furthermore, according to Table 4.1, our method demonstrates

Table 4.1: 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 4.2.

Setting	Accuracy (%)	Precision (%) Supported	Precision (%) Refuted	Precision (%) NEI	Recall (%) Supported	Recall (%) Refuted	Recall (%) NEI
Our w/ Text Evidence → DeBERTa V3	43.7	79.2	66.9	**33.9**	40.5	30.6	25.8
Our w/ Text and Image Evidence → DeBERTa V3	50.8	83.4	**69.3**	27.3	42.9	34.2	30.9
Our w/ Text Evidence → Llama 2	46.7	80.4	68.1	31.5	37.2	35.4	31.5
Our w/ Text and Image Evidence → Llama 2	**53.7**	**87.3**	60.3	32.4	**48.3**	**36.9**	**34.8**
Our w/Text Evidence → Stance detection model	40.5	79.4	68.3	32.3	43.7	34.2	14.4
Our w/ Text and Image Evidence → Stance detection model	45.4	76.8	66.9	34.7	35.7	36.2	30.9

strong precision performance for the "supported" label and strong recall performance in the "supported", "NEI", and "refuted" labels.

Table 4.2: Performance of claim verification in MOCHEG. DeBERTa V3, Llama-2-70b, and the stance detection model represent the fixed entailment models. Gold Evidence denotes ground truth text and image evidence while System Evidence means automatically retrieved text and image evidence.

Setting	**F-score (%)**
Our w/ Text Evidence → DeBERTa V3	42.7
Our w/ Text and Image Evidence → DeBERTa V3	45.1
Our w/ Text Evidence → Llama 2	43.9
Our w/ Text and Image Evidence → Llama 2	**48.2**
Our w/Text Evidence → Stance detection model	41.8
Our w/ Text and Image Evidence → Stance detection model	43.3
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**

Table 4.1 reveals that the best results are achieved when inputs incorporate both textual and image evidence. Perhaps unsurprisingly given its size, the zero-shot Llama-2-70b entailment surrogate model surpasses DeBERTa V3 in performance. Nevertheless, a notable issue persists, where the surrogate entailment models struggle to accurately deal with NEI claim labels.

Table 4.2 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 mod-

els. Although a difference remains in the result of human vs system prediction performance, our model surpasses the prior state-of-the-art system by 4.2% F-score.

4.2 Explanation Generation

Table 4.3: Performace of explanation generation. Our system outperforms MOCHEG on equivalent settings. Gold Evidence denotes ground truth text and image evidence while System Evidence means automatically retrieved text and image evidence. Gold Truthfulness denotes ground truth truthfulness label while System Truthfulness means the predicted truthfulness 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
Our w/ System Evidence, Gold Truthfulness	**36.7**	**17.9**	25.7	10.7	**87.3**
Our w/ System Evidence, System Truthfulness	34.3	16.8	25.4	10.4	87.1

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. [37], where we consider the input claim C, its truthfulness label Y_C, and the summary for fact-checking $\{T_1, T_2, ...\}$ generated from **MetaSumPerceiver**. 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 **MetaSumPerceiver** for this task. In the evaluation phase, we utilize the truthfulness label predicted by the fixed entailment models. Following Yao et al. [37], we utilize BART based to generate the ruling statement. Our evaluation metrics include ROUGE [94], BLEU [95], and Bertscore [96]. To assess the performance of explanation generation, we compare it with MOCHEG [37], as shown in Table 4.3.

Claim	Summary for fact-checking	Image evidence	Truthfulness
#1 On March 21, 2021, an adviser to former U.S. President Donald Trump said he was preparing to launch his own social media platform.	… Donald Trump has been banned from Twitter, Facebook, and Instagram since the Jan. 6 riot at the Capitol, but a senior advise to the former president says he plans to launch **his own social media site in the next few months** ...		**Supported**
#2 During the pandemic, I have been in D.C. voting regularly.	… Rep. Mark Pocan of Wisconsin says he's been " in DC voting regularly" since the COVID-19 pandemic hit Congress in May 2020, **but his office says that's a bit of a stretch**. ...		**NEI**
#3 Ex-AARP members wrote this letter criticizing the 'Biden Regime.	... The letter, originally from ex-AARP members in 2009, **criticized then-President Barack Obama's administration**. The post has been edited over the years to mislead readers, earning a rating of Mostly False. ...		**Refuted**
#4 By spring 2020, the sun had entered a lockdown period where its solar activity decreased to the point that famine, earthquakes, and freezing weather threatened life on Earth.	... Solar minimums are important because **they can affect how satellites orbit the Earth, as well as the intensity of cosmic rays and the brightness of sun's rays**....		**Refuted**

Figure 4.1: Evidence summary examples in the explanation generation task. The truthfulness column shows gold labels. For instance, the third claim's article primarily discusses a letter critiquing the Obama administration. However, given President Joe's past collaboration with Ex-President Obama, the letter was manipulated to criticize the 'Biden Regime.' This assertion lacks support from credible sources, making it a refuted claim.

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 gener-

ated 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.

We illustrate our generated summaries for fact-checking in Figure 4.1. Our results show that our summaries contain sufficient evidence to determine the accuracy of the claim label. Whether the truthfulness label is supported, labeled as NEI (No Evidence Identified), or refuted, we consistently provide evidence for a fact-checker to make a determination of the truthfulness of the claim.

Table 4.4: Evaluating the effectiveness of our multimodal multi-document claims using the pre-trained claim detection model.

Claim classes	**Category percentage(%)**
Unimportant Factual Sentence (UFS)	17.67
Check-worthy Factual Sentence (CFS)	68.6
Non-factual Sentence (NFS)	13.71

4.3 Multimodal Multi-document Claims Analysis

To assess the effectiveness of **KG2Claim**, we evaluate the generated claims with the check-worthy classification task. In Table 4.4, 68.6% of the sentences identified as check-worthy factual sentences (CFS) originate from our approach. This suggests that our generated claims encompass factual information that the general public would find interesting and worth verifying. Additionally, 17.67% of sentences in our dataset are factual but deemed unimportant

for fact-checking (UFS), indicating a lack of public interest in these claims. Finally, 13.71% of sentences in our data are non-factual (NFS), primarily comprising questions, beliefs, and declarations.

The results reveal two main points. Firstly, we effectively integrate multimodal multi-document sources, as over 60% of our generated claims focus on public interest. This indicates that our traversal algorithm is well-suited for knowledge graphs. Secondly, the content of our generated claims is closely tied to the US election in the ClaimBuster dataset [97], specifically in the policy domain rather than other areas.

Table 4.5: Performance of detecting truthfulness label in Multimodal Multi-document claims.

Truthfulness labels	**Category percentage(%)**
Entailment label	74.3
Neutral label	8.24
Contradiction label	17.46

4.4 Multimodal Multi-document Claims Truthfulness Label Test

In our assumption, we consider the multimodal multi-document claims to be entirely entailed with news articles, images, and videos, as these claims are generated from multimodal knowledge graphs. To verify this assumption, we tested the generated claims using our **MetaSumPerceiver** model. When the claim and multimodal multi-document sources are provided to the model, it determines whether the claim is entailed by the news using the entailment model. In Table 4.5, 74.3% of our claims are confirmed to be entailed with news articles, indicating a definite connection with the sources. Additionally, 8.24% of our

claims are categorized as neutral, meaning that the content does not contradict the news, but further consideration is needed to determine the accuracy of the relationships in these claims. Finally, 17.46% of our claims are contradiction claims, clearly indicating that these claims are unrelated to the news. The reason behind this is that the content of these claims revolves around advertisements in news articles.

The results provide insights that differ from our initial assumption. We suspect that the discrepancy may arise from the edge labels in the knowledge graphs. We believe it's essential to introduce additional edge labels to categorize less important information. This approach would help us steer clear of certain edges and nodes during the traversal of knowledge graphs.

Table 4.6: Accuracy for prompting entailment, neutral, contradiction claims with Llama-2-70b in Multi-News-Fact-Checking dataset.

Claims	**Accuracy(%)**
Entailment claims	78.3
Nentral claims	64.2
Contradiction claims	74.1

4.5 Refining Multi-News-Fact-Checking Dataset

In section 3.1.3, we described how we generated the claims and labels comprising our Multi-News-Fact-checking dataset. In total, our dataset consists of 1,687,200 claims and labels. However, in some cases, Llama-2-70b misunderstands the summary and predicts the wrong label for the claims. To assess the quality of our dataset, we employ Llama-2-70b once again for a thorough validation of our dataset's claims and respective labels (entailment, neutral, and contradiction). We provide the prompt we use for this double-checking procedure in the appendices A.1 and show the prompted claims in Figure A.1.

Table 4.7: Performance of claim verification in Multi-News-Fact-Checking dataset. We compare our method with other offline summarization models.

Setting	Accuracy (%)	Precision (%) Entailment	Precision (%) Contradiction	Precision (%) Neutral	Recall (%) Entailment	Recall (%) Contradiction	Recall (%) Neutral
PEGASUS → DeBERTa V3	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 → DeBERTa V3	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
Our → DeBERTa V3	36.7	**75.5**	28.9	27.5	41.0	21.7	**47.2**
Our (No RL) → Llama 2	42.6	41.0	**53.7**	34.6	54.8	37.8	29.6
Our → Llama 2	**45.6**	49.2	48.7	33.6	**56.9**	**44.1**	28.4

In Table 4.6, we show the performance of Llama-2-70b at predicting the label produced from the first phase of our dataset. We show that Llama-2-70b exhibits strong performance on entailment claims (acc 78.3%) and contradiction claims (acc 74.1%). We observe that Llama-2-70b performs worse at distinguishing neutral claims, registering an accuracy of only 64.2%. This is likely because the neutral category requires identifying that a specific piece of a claim is neither entailed or contradicted. Thus, this case is harder than either entailment or contradiction alone. We show examples of specific prompts and corresponding entailment, neutral, and contradictory claims.

Additionally, we discovered that most predictions from Llama-2-70b are similar to human predictions, especially in entailment and contradiction claims. To investigate this similarity further, we conducted a human test by randomly selecting 200 claims. The results for entailment claims revealed that 65% of them had the same prediction as Llama-2-70b. For contradiction claims, 73% of the sampled claims matched Llama-2-70b's predictions. Finally, in neutral claims, 62% of the sampled claims had the same prediction as Llama-2-70b.

4.6 Ablation

Additionally, we conducted ablation experiments for claim verification on our Multi-News-Fact-Checking dataset. A comparative analysis of our method with Llama-2-70b and other

offline summarization models, PEGASUS [98] and T5 large [99], is presented in Tables 4.8 and 4.7.

Similar to our results in MOCHEG, Tables 4.8 and 4.7 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.

We also established human performance upper bounds on our Multi-News-Fact-Checking dataset following MOCHEG's methodology. We randomly sampled 200 claims and assigned labels for their truthfulness based on gold evidence (the human written summaries from which the claims were generated), system evidence (our generated summaries), and no evidence, resulting in F-scores of 0.76, 0.65, and 0.23, respectively.

Table 4.8: Performance of claim verification in Multi-News-Fact-Checking dataset. DeBERTa V3 and Llama-2-70b serve as the fixed entailment models. Gold Evidence refers to claim labels based on gold standards, whereas System Evidence indicates our predicted claim labels.

Setting	**F-score (%)**
Our w/ DeBERTa V3	39.9
Our w/ Llama 2	**43.4**
Our w/ Llama 2(No RL)	41.8
PEGASUS w/ DeBERTa V3	25.4
PEGASUS w/ Llama 2	**30.8**
T5 large w/ DeBERTa V3	28.5
T5 large w/ Llama 2	**32.7**
Human w/o Evidence	23.0
Human w/ System Evidence	65.0
Human w/ Gold Evidence	**76.0**

Chapter 5

Discussion

Given the societal importance of fact-checking applications, it is important that the limitations of our methods 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 **MetaSumPerceiver** during training. One potential solution is using a textual entailment model adept at managing this uncertainty or excluding such instances during training. Furthermore, the experimental outcomes from the **KG2Claim** method reveal that 30% of the generated claims are not entailed with the news sources. The challenge lies in the possibility that multimodal multi-document knowledge graphs might incorporate irrelevant information. A potential remedy is to diversify the set of edge labels in the knowledge graphs. We propose incorporating additional labels, such as the content label, ads label, and quote label. This approach would help prioritize which edges are more crucial for traversal. Lastly, 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.

MetaSumPerceiver, 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.

Chapter 6

Conclusions

We introduce **MetaSumPerceiver**, a summarization model designed to produce concise, informative summaries for claim fact-checking from complex multimodal datasets. Our model's flexible architecture can accommodate arbitrary numbers of documents and types of inputs, including documents, images, and claims by leveraging a perceiver-based architecture. In addition, we propose **KG2Claim**, a text generation pipeline to produce the claims from the knowledge graphs. Our text generation approach can generate claims related to multimodal mult-document information.

We train our model using a novel reinforcement learning approach in order to generate summaries useful for verifying the truthfulness of claims. Our experimental assessments on the MOCHEG and our Multi-News-Fact-Checking datasets highlight **MetaSumPerceiver**'s robust performance in claim verification tasks and demonstrate its effectiveness in real-world fact-checking scenarios. This contribution underscores **MetaSumPerceiver**'s potential to streamline fact-checking processes in today's multimodal information landscape. Moreover, we release the publicly accessible Multi-News-Fact-Checking dataset, aimed at assisting researchers in developing multi-document fact-checking methods.

Furthermore, we employ our text generation pipeline to produce claims in the NewsStories dataset. Our analysis of the generated claims reveals that more than 60% are factual, drawing public interest in fact-checking. Additionally, testing these claims with **MetaSumPerceiver** demonstrates that over 70% of them are entailed with the news sources. According to the

above experiment, the conclusion indicates that more than 60% of the claims are entailment claims, and people wish to discern whether they are correct.

Bibliography

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

[2] B Borel, K Sheikh, F Husain, A Junger, E Biba, D Blum, and B Urcuioli. The state of fact-checking in science journalism. *Cambridge, MA: Massachusetts Institute of Technology Knight Science Journalism Program*, 2018.

[3] 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. Llama 2: Open foundation and fine-tuned chat models, 2023.

[4] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120

(11), mar 2023. doi: 10.1073/pnas.2208839120. URL https://doi.org/10.1073%2Fpnas.2208839120.

[5] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, apr 2023. doi: 10.1145/3544548.3581196. URL https://doi.org/10.1145%2F3544548.3581196.

[6] Sarah Kreps, R. Miles McCain, and Miles Brundage. 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, 2022. doi: 10.1017/XPS.2020.37.

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

[8] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. Ai model gpt-3 (dis)informs us better than humans, 2023.

[9] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective, 2017.

[10] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. doi: 10.1126/science.aap9559. URL https://www.science.org/doi/abs/10.1126/science.aap9559.

[11] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020. doi: 10.1109/JSTSP.2020.2987728.

[12] Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, 60(2):103219, mar 2023. doi: 10.1016/j.ipm.2022.103219. URL https://doi.org/10.1016%2Fj.ipm.2022.103219.

[13] Andrea Ceron and Paride Carrara. Fact-checking, reputation, and political falsehoods in italy and the united states. *New Media & Society*, 25(3):540–558, 2023. doi: 10.1177/14614448211012377. URL https://doi.org/10.1177/14614448211012377.

[14] Nicolas Berlinski, Margaret Doyle, Andrew M. Guess, Gabrielle Levy, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, and Jason Reifler. The effects of unsubstantiated claims of voter fraud on confidence in elections. *Journal of Experimental Political Science*, 10(1):34–49, 2023. doi: 10.1017/XPS.2021.18.

[15] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):1–29, 03 2016. doi: 10.1371/journal.pone.0150989. URL https://doi.org/10.1371/journal.pone.0150989.

[16] Brian Chen, Xudong Lin, Christopher Thomas, Manling Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. Joint multimedia event extraction from video and article, 2021.

[17] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction, 2020.

[18] Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Dan Napierski, and Marjorie Freedman. Gaia: A fine-grained multimedia knowledge extraction system. In *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[19] Sha Li, Heng Ji, and Jiawei Han. Document-level event argument extraction by conditional generation, 2021.

[20] Jian Liu, Yufeng Chen, and Jinan Xu. Multimedia event extraction from news with a unified contrastive learning framework. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1945–1953, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548132. URL https://doi.org/10.1145/3503161.3548132.

[21] Aman Khullar and Udit Arora. Mast: Multimodal abstractive summarization with trimodal hierarchical attention, 2020.

[22] Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. Long text and multi-table summarization: Dataset and method, 2023.

[23] Michal Rott and Petr Červa. Speech-to-text summarization using automatic phrase extraction from recognized text. volume 9924, pages 101–108, 09 2016. ISBN 978-3-319-45509-9. doi: 10.1007/978-3-319-45510-5_12.

[24] Shagan Sah, Sourabh Kulhare, Allison Gray, Subhashini Venugopalan, Emily Prud'Hommeaux, and Raymond Ptucha. Semantic text summarization of long videos. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 989–997, 2017. doi: 10.1109/WACV.2017.115.

[25] Clément Rebuffel, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari. A hierarchical model for data-to-text generation, 2019.

[26] Ratish Puduppully and Mirella Lapata. Data-to-text Generation with Macro Planning. *Transactions of the Association for Computational Linguistics*, 9:510–527, May 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00381. URL https:

//doi.org/10.1162/tacl_a_00381. _eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00381/1924176/tacl_a_00381.pdf.

[27] Fei Wang, Zhewei Xu, Pedro Szekely, and Muhao Chen. 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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.371. URL https://aclanthology.org/2022.naacl-main.371.

[28] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and Joāo Carreira. Perceiver io: A general architecture for structured inputs & outputs, 2022.

[29] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and Joāo Carreira. Perceiver io: A general architecture for structured inputs & outputs, 2022.

[30] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[31] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. pages 177–190, 01 2005. ISBN 978-3-540-33427-9. doi: 10.1007/11736790_9.

[32] Adam Poliak. A survey on recognizing textual entailment as an nlp evaluation, 2020.

[33] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

[34] Shon Otmazgin, Arie Cattan, and Yoav Goldberg. F-coref: Fast, accurate and easy to use coreference resolution, 2022.

[35] Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. Realistic evaluation principles for cross-document coreference resolution, 2021.

[36] Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.225. URL https://aclanthology.org/2021.findings-emnlp.225.

[37] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 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, jul 2023. doi: 10.1145/3539618.3591879. URL https://doi.org/10.1145%2F3539618.3591879.

[38] Sahil Garg, Greg Ver Steeg, and Aram Galstyan. Stochastic learning of nonstationary kernels for natural language modeling, 2018.

[39] Miguel Ballesteros and Yaser Al-Onaizan. Amr parsing using stack-lstms, 2017.

[40] Chunchuan Lyu and Ivan Titov. Amr parsing as graph prediction with latent alignment, 2018.

[41] Hanwen Zheng, Sijia Wang, and Lifu Huang. A survey of document-level information extraction, 2023.

[42] Amelia Devi Putri Ariyanto, Chastine Fatichah, and Diana Purwitasari. Semantic role labeling for information extraction on indonesian texts: A literature review. In *2023 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 119–124, 2023. doi: 10.1109/ISITIA59021.2023.10221008.

[43] Matt Grenander, Shay B. Cohen, and Mark Steedman. Sentence-incremental neural coreference resolution, 2023.

[44] Bernd Bohnet, Chris Alberti, and Michael Collins. Coreference Resolution through a seq2seq Transition-Based System. *Transactions of the Association for Computational Linguistics*, 11:212–226, 03 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00543. URL https://doi.org/10.1162/tacl_a_00543.

[45] Shan Jie Yong, Kuicai Dong, and Aixin Sun. Docor: Document-level openie with coreference resolution. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1204–1207, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3573038. URL https://doi.org/10.1145/3539597.3573038.

[46] Nathanon Theptakob, Thititorn Seneewong Na Ayutthaya, Chanatip Saetia, Tawunrat Chalothorn, and Pakpoom Buabthong. A cross-document coreference resolution approach to low-resource languages. In *Knowledge Science, Engineering and Management: 16th International Conference, KSEM 2023, Guangzhou, China, August*

16–18, 2023, Proceedings, Part II, page 422–431, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-40285-2. doi: 10.1007/978-3-031-40286-9_34. URL https://doi.org/10.1007/978-3-031-40286-9_34.

[47] Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. A brief survey on recent advances in coreference resolution. *Artif. Intell. Rev.*, 56(12):14439–14481, December 2023.

[48] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans, 2020.

[49] Yao Yao, Zuchao Li, and Hai Zhao. Learning event-aware measures for event coreference resolution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13542–13556, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.855. URL https://aclanthology.org/2023.findings-acl.855.

[50] Tong Shen, Fu Zhang, and Jingwei Cheng. A comprehensive overview of knowledge graph completion. *Knowledge-Based Systems*, 255:109597, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.109597. URL https://www.sciencedirect.com/science/article/pii/S095070512200805X.

[51] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015. doi: 10.1609/aaai.v29i1.9491. URL https://ojs.aaai.org/index.php/AAAI/article/view/9491.

[52] Baoxu Shi and Tim Weninger. Open-world knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11535. URL https://ojs.aaai.org/index.php/AAAI/article/view/11535.

[53] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.

[54] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

[55] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, page 1112–1119. AAAI Press, 2014.

[56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[57] Liang Yao, Chengsheng Mao, and Yuan Luo. Kg-bert: Bert for knowledge graph completion, 2019.

[58] Yichi Zhang, Zhuo Chen, Wen Zhang, and Huajun Chen. Making large language models perform better in knowledge graph completion, 2023.

[59] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

[60] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention, 2021.

[61] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

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

[63] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, and Jifeng Dai. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2691–2700, June 2023.

[64] Zhu Junnan, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. Multimodal summarization with guidance of multimodal reference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9749–9756, 04 2020. doi: 10.1609/aaai.v34i05.6525.

[65] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1448. URL https://aclanthology.org/D18-1448.

[66] Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, Florence, Italy, July

2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1659. URL https://aclanthology.org/P19-1659.

[67] 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. Llama: Open and efficient foundation language models, 2023.

[68] 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. Palm-e: An embodied multimodal language model, 2023.

[69] OpenAI. Gpt-4 technical report, 2023.

[70] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 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, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

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

[72] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[73] 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. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

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

[75] 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. Secrets of rlhf in large language models part i: Ppo, 2023.

[76] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020.

[77] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, 2021.

[78] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey,

M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2022.

[79] Zhenchang Xing, Qing Huang, Yu Cheng, Liming Zhu, Qinghua Lu, and Xiwei Xu. Prompt sapper: Llm-empowered software engineering infrastructure for ai-native services, 2023.

[80] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023.

[81] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models, 2023.

[82] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[83] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[84] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[85] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[86] Adina Williams, Nikita Nangia, and Samuel Bowman. 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, 2018. URL http://aclweb.org/anthology/N18-1101.

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

[88] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441. URL https://aclanthology.org/2020.acl-main.441.

[89] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2023.

[90] Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model, 2019.

[91] Yuval Kirstain, Ori Ram, and Omer Levy. Coreference resolution without span representations. 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 14–19, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.3. URL https://aclanthology.org/2021.acl-short.3.

[92] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[93] Xuefeng Bai, Yulong Chen, and Yue Zhang. Graph pre-training for amr parsing and generation, 2022.

[94] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[95] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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,

2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

[96] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[97] Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. A benchmark dataset of check-worthy factual claims, 2020.

[98] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020.

[99] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Appendices

Appendix A

Prompting Design

A.1 Multi-document Generated and Checked Prompts

In the chapter 3.1.3 and 4.5, I reference the prompts for generating and checking in the provided information:

- **Entailment claim**: 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.).

- **Neutral claim**: 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.

- **Contradiction claim**: 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

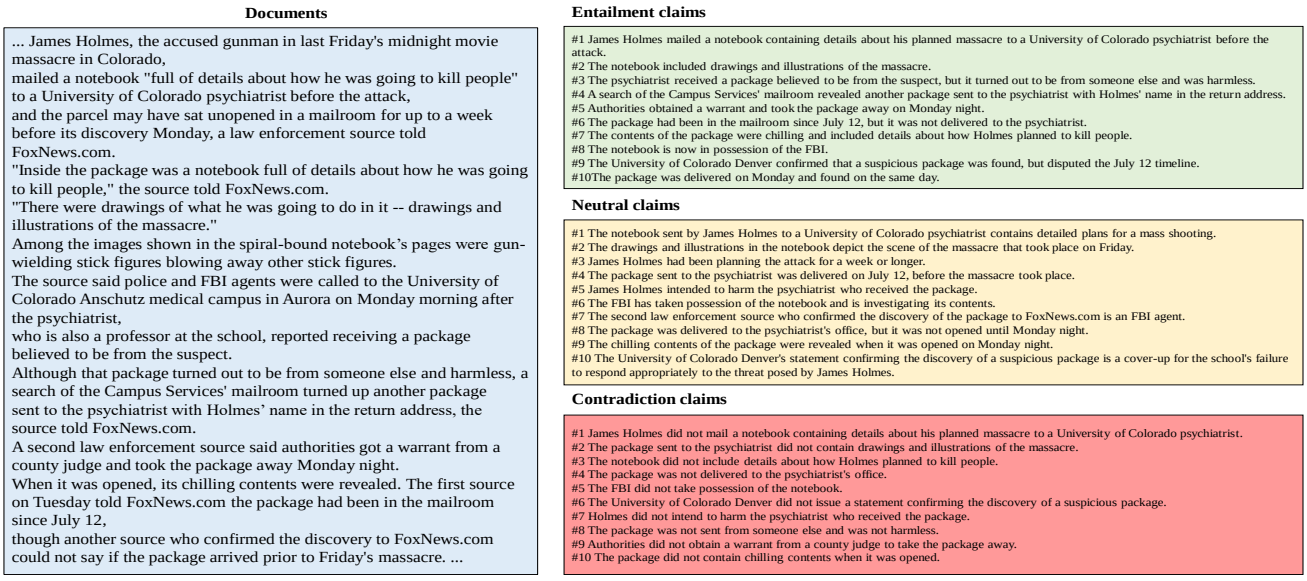

Figure A.1: The prompted entailment, neutral, contradiction claims from Llama-2-70b.

statements like ”may” or ”could” - your claim should state things as a contradiction fact.

- **Double check claim**: 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.