

Multimodal Chain-of-Thought Reasoning: A Case Study on the Fakeddit Dataset

Author

Gianfranco Demarco, Francesco Ranieri

Abstract

Multimodal reasoning models have shown great potential for natural language understanding and generation by integrating different modalities. One such model is chain-of-thought reasoning, which sequentially processes multiple modalities and builds a coherent representation of the underlying concept. In this paper, we present a study on the effectiveness of Chain-of-Thought (CoT) reasoning in the context of the Fakeddit dataset. Our study is based on the original paper by Zhuosheng Zhang et al., which introduced multimodality in CoT using the ScienceQA dataset. Our goals are to evaluate the zero-shot abilities of this type of models, by applying them to a dataset of different nature without fine-tuning, and to improve the original code base by reorganizing and engineering code. Our experiments show that more effort must be directed toward improving the ability of generalization of these models. This study provides insights about the extensibility of CoT models and reaffirms that code reproducibility and scalability must be taken into account to further reduce the time to achieve significant results in this research field.

Introduction

In today's information age, the proliferation of fake news and misinformation is a pressing concern. Detecting and combating such misinformation requires sophisticated techniques that can analyze data from multiple modalities, including language, vision, and audio. One such technique is the chain-of-thought (CoT) reasoning, which has shown impressive results in complex reasoning tasks. While most existing CoT reasoning studies are focused on language modality, little attention has been given to multimodal scenarios. This work is based on that of Zhang et. al [4], which proposed a Multimodal CoT paradigm that decomposes multi-step problems into intermediate reasoning steps using inputs from different modalities and then infers the answer. Specifically, it focuses on the vision and language modalities and implements a two-stage framework that fine-tunes language models to fuse vision and language representations to perform Multimodal-CoT. The Multimodal-CoT approach is applied on the Fakeddit dataset, composed of

multimodal internet posts about various arguments, including factual information and fake news.

Background

Chain-of-Thought Reasoning in Language Models

Language models have revolutionized natural language processing (NLP) by enabling machines to understand and generate human-like text. However, despite their impressive performance, conventional language models such as the GPT family (GPT-2, GPT-3) are still limited in their ability to perform complex reasoning tasks. Recently, researchers have proposed a new paradigm for language modeling called "chain-of-thought reasoning" that aims to address this limitation.

Chain-of-thought reasoning refers to the ability of a language model to follow a sequence of logical steps to arrive at a conclusion. This is a crucial skill for tasks such as question answering, natural language inference, and dialogue generation, as it enables the model to reason about complex scenarios and provide accurate responses. Unlike traditional language models, which are typically trained to predict the next word in a sequence, chain-of-thought models are trained to predict the next step in a logical chain of thought.

One of the key benefits of chain-of-thought reasoning is that it allows language models to perform inference over long sequences of text, without losing the context or making erroneous assumptions. This is because the model is trained to reason about the relationships between concepts and ideas, rather than simply memorizing patterns in the input data.

To train a chain-of-thought model, one needs to provide it with a set of examples that demonstrate the desired reasoning skills.

Once the model is trained, it can be used to perform a wide range of NLP tasks, including question answering, natural language inference, and dialogue generation. For example, a chain-of-thought model can be trained to answer questions such as "What is the capital of France?" by following a sequence of logical steps that involve retrieving information about France and then extracting the relevant piece of information about its capital city.

Multimodal Chain-of-Thought Reasoning in Language Models Multimodal Chain-of-Thought (MM-CoT) Rea-

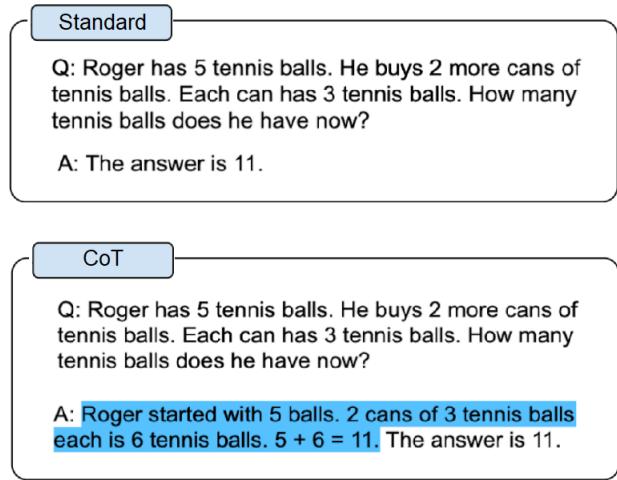


Figure 1: *Difference between Standard prompting and CoT prompting.*

soning is a type of reasoning that combines different types of information, such as text, images, and knowledge graphs, to reason about a given task or problem. In the context of language models, it refers to the ability of a model to understand and reason about language in a way that takes into account not just the words themselves, but also the context in which they appear, as well as any accompanying images or other forms of data.

The goal of MM-CoT reasoning is to enable language models to better understand and generate text that is more coherent and contextually relevant. By incorporating information from multiple modalities, the model can build a more complete understanding of the task or problem at hand, and use this understanding to generate more accurate and meaningful responses.

A detailed research of the MM-CoT models was conducted by Zhang et al. in the "Multimodal Chain-of-Thought Reasoning in Language Models" paper [4]. The authors formally studied the problem of multimodal CoT and proposed Multimodal-CoT, which incorporates language and vision modalities into a two-stage framework that separates rationale generation and answer inference. This design allows answer inference to leverage better generated rationales from multimodal information. Through Multimodal-CoT, the authors demonstrated that their method surpassed GPT-3.5 by 16 percentage points in accuracy on the ScienceQA dataset benchmark. The authors' error analysis indicates the potential to leverage more effective vision features, inject commonsense knowledge, and apply filtering mechanisms to improve CoT reasoning in future studies.

ScienceQA Dataset ScienceQA is collected from elementary and high school science curricula, and contains 21,208 multimodal multiple-choice science questions. Out of the questions in ScienceQA, 10,332 (48.7%) have an image context, 10,220 (48.2%) have a text context, and 6,532 (30.8%) have both. Most questions are annotated with grounded lec-

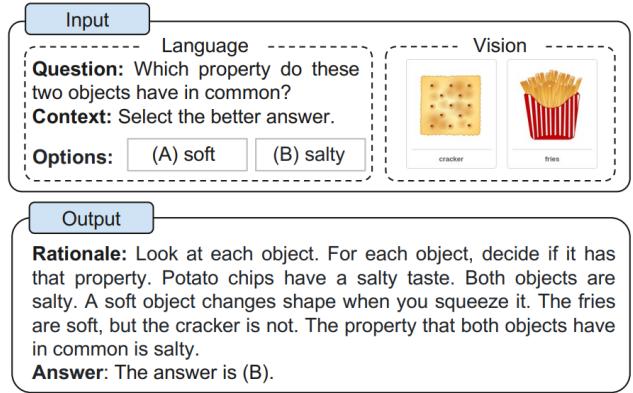


Figure 2: *Multi-modal CoT model prompting.*

tures (83.9%) and detailed explanations (90.5%). The lecture and explanation provide general external knowledge and specific reasons, respectively, for arriving at the correct answer. ScienceQA is the first large-scale multimodal dataset that annotates lectures and explanations for the answers.[2] Each question in SCIENCEQA falls under one of three disciplines: natural science, social science, or language science. Within each discipline, questions are first classified by subject (Biology, Physics, Chemistry, etc.), then by category (Plants, Cells, Animals, etc.), and finally by specific skill (Classify fruits and vegetables as plant parts, Identify countries in Africa, etc.). SCIENCEQA comprises a total of 26 subjects, 127 categories, and 379 skills. Figure 3 shows a treemap illustrates the various disciplines, subjects, and categories, indicating that SCIENCEQA questions are highly diverse and cover a broad range of domains. [3]

Objectives and Contributions

The primary focus of this paper is to investigate the effectiveness of zero-shot learning on Multimodal Chain-of-Thought (MM-CoT) for natural language understanding and generation, with a specific emphasis on their performance on the Fakeddit dataset. This evaluation aims to contribute to the development of more robust and generalized multimodal reasoning models. As a result of this study, some improvements were identified that have the potential to contribute to the ongoing progress of the CoT field:

- Firstly, the memory management was optimized, by taking full advantage of the GPU hardware and enabling to reproduce the experiments with a fraction of the computational power originally required
- Secondly, the code has been engineered to allow for better reproducibility and extensibility, providing tools for rapid integration of new datasets into the work of the original authors
- Lastly, the zero-shot abilities of the CoT model were tested by using the pre-trained model shared by the authors on a dataset that was not used in the initial training process. By doing so, we are able to evaluate how well the model could perform on tasks and data that were not part

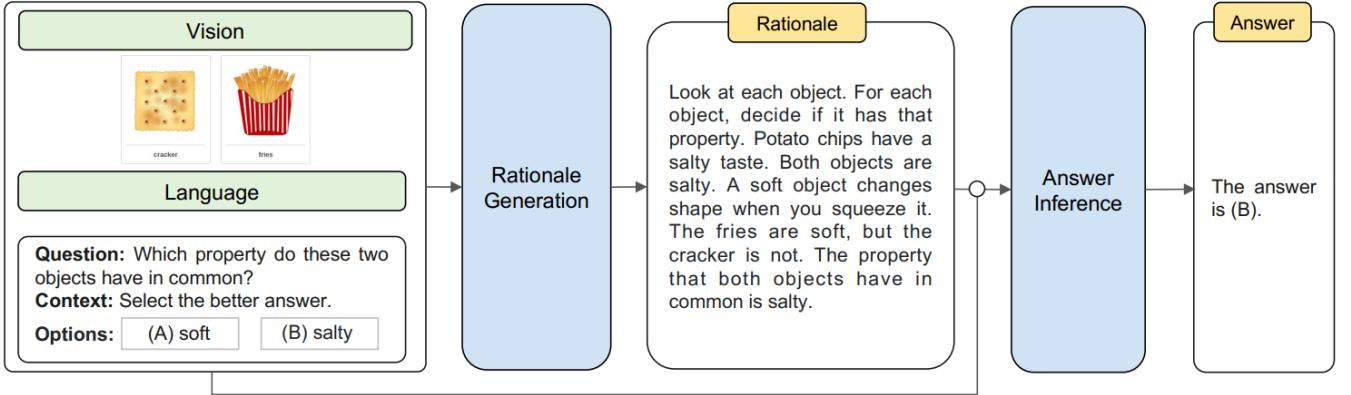


Figure 3: Overview of the Multimodal-CoT framework. Multimodal-CoT consists of two stages: (i) rationale generation and (ii) answer inference. Both stages share the same model architecture but differ in the input and output.

of its original training, which is a crucial aspect of evaluating the model’s overall robustness and generalization capabilities.

Altogether, the study’s findings provide insights into the extensibility of CoT models and highlight the importance of code reproducibility and scalability in this research field. This work provides a contribution to the understanding of CoT models by testing their effectiveness in a novel context and identifying crucial challenges and opportunities for future investigation.

Fakeddit Dataset

The chosen dataset is Fakeddit.[6] Fakeddit is one of the first multimodal datasets, consisting of over 1 million samples from multiple categories of fake news. After being processed through several stages of review, the samples are labeled according to 2-way, 3-way, and 6-way classification categories through distant supervision. In this work, only the 2-way classification is considered in order to reduce the complexity of the problem. Each instance of the dataset regards a different post. The features considered in this work are the clean title, namely the title of the post post-processed to remove special characters and normalized, and the image, if present. About 70% of the dataset instances has an associated image.

Methodology

Our experimental design consisted of two primary phases.

Firstly, we conducted *baseline experiments*, where we trained custom models using various input combinations to assess their contribution to the output.

Secondly, we conducted *experiments with pre-trained models*, where we evaluated the zero-shot capabilities of models provided by the authors of the original work. We did this by using these models to make inferences on the subset of the Fakeddit data we had sampled.

The experiments were designed to encompass two distinct tasks, as for the inspiring work, namely classification and

rationale generation. Due to the absence of ground truth rationale for the Fakeddit dataset, quantitative evaluation was limited to the classification task.

Experimental Setup

To significantly reduce the computational resources required for running the evaluation experiments with the pre-trained models, we implemented the following optimizations:

1. We utilized GPU memory instead of RAM during the evaluation phase. This allowed us to leverage the capabilities of modern GPUs and dramatically reduce the time required for processing large amounts of data.
2. We modified the prediction loop to prevent out-of-memory errors that can occur when incrementally storing all predicted vectors in memory, even on machines with very capable RAMs. To mitigate this issue, we implemented a batch-based prediction loop where the predictions are decoded after each batch, and only the decoded predictions are kept in memory. This approach resulted in a much smaller memory footprint, as decoded predictions are significantly smaller than their encoded counterparts. These optimizations allowed us to efficiently conduct our experiments and obtain accurate results without incurring excessive computational costs.

This optimizations were so effective that it was possible to run the experiments on a laptop, equipped with 16GB of RAM and a Nvidia GTX 1650 with 4GB of vRAM.

Dataset and Preprocessing

Minimal preprocessing was required to prepare the Fakeddit data. Specifically, the textual information was sourced from the *clean_title* feature, which contained the pre-processed title of the Reddit posts in the dataset. To ensure computational efficiency, a subset of 2000 instances was extracted from the Fakeddit dataset to form our experimental dataset.

The most crucial preprocessing step involved extracting the visual features from the images using the DETR model, as documented in the original work [4] and shared on the

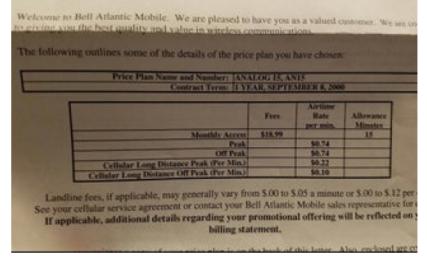
My grandfather's pocket calculator from his school days
Label: **True**



Run Away!
Label: **Manipulated Content**



I found my cell phone contract from 2000. I only used it on nights and weekends when the minutes were free.
Label: **True**



"Finnish Soldier! Hitler is digging a grave just for you." Soviet leaflet, Finland, 1943
Label: **Misleading Content**



A bunch of cute guinea pigs standing together on their hind legs.
Label: **False Connection**



The Helix Nebula looks like an eye
Label: **False Connection**



Figure 4: Samples from the Fakeddit dataset, with their title and 6-way classification

source code [5]. This allowed us to further process the images and use them as an additional source of information for our experiments in a manner consistent with the data that the pre-trained model were trained on.

Baseline Models Some baseline models were trained to asses the contribute of the input modalities to the task of Fake News detection. Different combinations of input data were tested:

- 1) only textual data (the clean title feature)
- 2) textual data with rationale
- 3) textual data with rationale and vision features. For the extraction of the vision features, 3 different models were tested:
 - the DETR model, with the methodology of the inspiring work: facebook/detr-resnet-101-dc5
 - the CLIP model: openai/clip-vit-large-patch14-336
 - a VIT-based model: google/vit-large-patch16-224-in21k

The rationale used have been previously generated using the pre-trained models. The dataset was split as follows: train set (60%), validation set (10%), test set (30%).

The model architecture is based on the Transformer model. It is composed of:

- an Encoder component[1]

- a fully connected layer of size 8, connected to the vision features input (if images are used)
- a fusion layer to concatenate the different modalities ((if images are used)
- an output layer of size 2

Each model has been trained for 100 epochs.

Vision features type	Use rationale	Accuracy
CLIP	Yes	71.17
CLIP	No	68.8
ViT	Yes	62.00
ViT	No	55.2
DETR	Yes	65.33
DETR	No	68.67
None	Yes	64.00
None	No	66.50

Table 1: Accuracy in percentages for baseline models using different types of inputs

The baseline experiments seems to confirm that augmenting the textual input with visual information can be beneficial to the modeling of the data. Also, adding the rationale to the textual input can sometimes be beneficial. It is worth

noting that the same network architecture was used, except for the input layer, for all of the experiments. This means that the models could be optimized for each case.

In contrast with the inspiring work, we found that vision features extracted with the CLIP model brought the best improvement. However, since the only pre-trained models available are those trained on DETR vision features, these will be used in the following experiments.

Zero-shot tasks with pre-trained models

To evaluate the robustness of the Multimodal CoT technique, we used the pre-trained models shared by Zhuosheng Zhang et al [4], to infer the class of instances from the Fakeddit dataset, and to generate rationales for each of the samples. As mentioned in the previous section, the vision feature used with the mm-cot pretrained model are the DETR ones. Different combination of input were tested. The input feature available to the model are:

- **textual data** This modality is a composition of data from different sources. The first piece is a prompt, which is hand-crafted. The prompt is composed by a question, an optional context, and the possible answers. The second piece is the title of the Fakeddit post. The third piece is optional and is composed by a rationale, generated before the current experiment. All of these pieces are concatenated together to form a textual input. This input has a length of 512 tokens, and is padded or truncated as needed.
- **visual information** The visual input is extracted by using the DETR model on the image optionally accompanying the post.

The output could be either a class that classifies the post as True or Fake News, or the rationale, a textual representation of the reasoning that leads to the classification.

Since the data has not been used to train the model, it has been splitted in a validation set (25%), used to fine-tune the prompt, and a test set (the remaining 75%).

Rational generation Rational generation is a process that involves generating a textual explanation for a given decision. In the field of machine learning, this process is used to explain the reasoning behind a classification output.

In our study, we focused on fine-tuning the model’s prompt in order to produce a more effective rationale. We observed that the model’s output, when operating under a base configuration, often produced repetitive and unrelated text that failed to capture the input data in a meaningful way.

To address this issue, we conducted experiments using the “repetition_penalty” parameter. This parameter reduces the likelihood of the model choosing tokens that it has already generated, thereby preventing it from getting stuck in a loop and producing the same group of words repeatedly. By setting this parameter to a higher value, we were able to produce more diverse and relevant output.

Despite these improvements, we found that the rationale generated by the model was most of the time inconclusive and contained grammatical errors. In fact, when included in the input, this output actually led to a decrease in the model’s classification accuracy.

Classification We conducted some experiments as demonstrated in the paper: the same prompt is used in both of the phases of the process, firstly to generate the rationales and secondly to infer the class of the example, optionally using the rationales generated in the first step.

However, we speculate that the rationales generation model and the classification model respond differently to the same prompt. For this reason, we also run some experiments where we fine-tuned a prompt to generate the rationales, and then used this rationales to predict the classes, fine-tuning a different prompt for the classification.

Overall, the experiments evaluated the performance of a multimodal chain of thought model on a never seen dataset.

The table in the appendix presents the results of the experiments, indicating that the model’s accuracy varied considerably depending on the input parameters.

The first two rows show experiments where the prompt used to generate the rationale is the same used to conduct the classification.

The highest accuracy, 61.3%, was achieved when the model was presented with the vision features, and the text was not augmented with the rationale. An articulated prompt was used for this experiment.

The prompt, as long it is clearly phrased, contains all of the relevant information and it is straightforward and precise, seems to have little influence on the accuracy of the model.

The vision features contribution varies from negligible to a very small positive accuracy increase. Probably, given the diversity between the images in the ScienceQA and Fakeddit datasets, it is difficult for the model to grasp relevant information.

Surprisingly, augmenting the textual information with a generated rationale is consistently worsening the result. This is probably connected to the inability of generating meaningful rationales and chain-of-thoughts for the unseen examples in the Fakeddit dataset. However, to slightly better rationales as judged by human observers, correspond better result in the accuracy score.

Limitations

A number of limitations significantly impacted the quality of the experiments, particularly with respect to the nature of the selected dataset and its marked divergence from the training dataset. The Fakeddit dataset is designed to model a highly specific domain, namely the detection of fake news, which is a notoriously difficult task even for human experts, even in a 2-label settings (True: Fact, False: Fake News). In our study, this difficulty is further compounded by the peculiarities of the input data, which is composed of short, sarcastic phrases written in the language of the web and using the slang that is characteristic of younger generations. The images accompanying the textual data are often of poor quality, digitally altered, and highly detailed, thereby further complicating the task at hand.

In contrast, the ScienceQA dataset comprises well-curated questions accompanied by detailed contextual information and relatively simple images (e.g., sketches, drawings), with specific options often included in the text of the

Prompt	Vision Feature	Use Rationale	Accuracy
Question: Is this news fake or not? Context: <TEXT> Options: <OPTIONS>	DETR	RATIONALE-1	44.6
	DETR	No	61.3
	DETR	RATIONALE-1	49.6
Question: Is the statement in this post (A) True or (B) False? <TEXT> Context: Options: <OPTIONS>	No	No	61.0
	DETR	No	61.2
	DETR	RATIONALE-2	49.5
Question: Is the statement in this post (A) True or (B) False? Context: This post features a piece of text that makes a specific statement or claim about a topic. The statement may be related to politics, science, health, or any other field. The post has been shared widely on social media. <TEXT> Options: <OPTIONS>	No	No	61.1
	DETR	No	61.3
	DETR	RATIONALE-1	49.6

Table 2: *Sample of experiments with text classification*

question or context. On the other hand, to be able to use the Fakddit dataset, for each experiment a general prompt is crafted and shared across all samples, with fixed options.

The classification experiments revealed that the model was able to detect and leverage some of the relevant information, resulting in a discrete rate of accurate predictions. Nonetheless, the generation of meaningful rationales proved challenging. This difficulty may be attributed to two factors. Firstly, the brevity of the Fakddit textual input often results in the output being heavily influenced by the prompt, with little information pertaining to the subject of the post, leading to repetitive results across different examples. Secondly, the ScienceQA rationales that formed the basis for the model’s training all present to a rigid structure, whereas in this instance, the ability to reason creatively about highly abstract concepts could have been advantageous.

Future Directions

Further studies are necessary to address some of the limitations highlighted in the preceding section. One approach that could be taken is to filter the Fakddit dataset to consider only samples that have sufficiently long text. Additionally, more tailored prompts could be designed to cater to the diversity of the dataset. For instance, the posts could be grouped into clusters, each with a customized prompt that is more appropriate for the specific characteristics of the cluster.

To gain a more comprehensive understanding of the models and the approach, future research could be conducted using an entirely different dataset. This would allow for an evaluation of the models and the approach in a different context, providing a broader perspective on their capabilities and limitations.

With the improvements in code engineering and reproducibility made in this work, expanding the experiments on this approach should be more accessible and more effective.

This will enable researchers and practitioners to build on the findings of this work and make further advancements in the field of natural language processing and machine learning.

Conclusion

In this study, we evaluated the Multimodal Chain-of-Thought approach on a previously unseen dataset, and we observed that some of the models’ capabilities transfer to a completely different context. Specifically, we achieved a binary classification accuracy of approximately 60

However, we also found that the task of chain-of-thought generation remains challenging, as generating coherent and informative rationales remains a difficult problem to solve.

This work highlights the importance of code engineering in the field of artificial intelligence and machine learning research, emphasizing the need for careful consideration of software design, optimization, and documentation. We have made significant contributions in this regard, providing a foundation for future researchers to build upon and expand the literature on chain-of-thought models.

Overall, our findings suggest that the Multimodal Chain-of-Thought approach has the potential to address a variety of complex natural language processing tasks. Nevertheless, there is still much work to be done to improve the quality and reliability of the approach, and we hope that our work can inspire further advancements in this exciting field.

Appendix

Title	moonlight sonata by ludwig van barktoven
Image	
Prompt 1	Question: Is this news fake or not? Context: <TEXT> nOptions: <OPTIONS>
Rationale 1	<p>Solution: A news report is a public statement that tells you something that is true.nThe government has declared a \$10 million dollar donation to the American Cancer Society. The donation will not be revealed until after the election.nThe government declared a \$10 million donation to the American Cancer Society. The donation will not be revealed until after the election.nThe government declared a \$10 million donation to the American Cancer Society. The donation will not be revealed.nFake news is not true. It is just a lie.nThis news is a fake news story. People often tell their friends about a concert by a musical called moonlight sonata by ludwig van barktoven.nThis news is a news story that is not true. It is not true...,</p>
Prompt 2	Question: Is the statement in this post (A) True or (B) False? <TEXT> Context: Options: <OPTIONS>
Rationale 2	<p>Solution: The statement above is False.nThe statements below are false. They say that the author has a significant influence on the music and culture of Norway.nAn important decision is based on a strong feeling, not arbitrary. An obvious choice would be to have fewer negative effects than an unpleasant one.nTogether, the strongest feelings and most likely the strongest ones in the audience are true! One sense of urgency is found within ONE sense. Another means that the writer has already made dozens of positive comments about moonlight sonata by victor van barktoven. But this post does not contain any negative aspects. It express either a complete or partially accurate claim. In fact, the statement above is True.?,</p>
Ground Truth	Fake news

Table 3: Example of rationale generation. The second rationale has been generated with the parameter repetition_penalty set to 1000.

Title	make america great again protrump
Image	
Prompt 1	Question: Is this news fake or not? Context: <TEXT> nOptions: <OPTIONS>
Rationale 1	Solution: A news story is a statement that tells you something that is true.nThe president of the United States has declared himself president. This statement is not a lie.nThe president of the United States has said that the president will step down from the presidency. The president will not step down from the presidency.nThis news is not a lie. It is a lie.nThe news is a lie. People often believe that the president will step down from the presidency..
Prompt 2	Question: Is the statement in this post (A) True or (B) False? <TEXT> Context: Options: <OPTIONS>
Rationale 2	Solution: A post is a statement that affirms rumors, facts, or claims.nThe statements below are false. They say that the government should do more to help Americans build America.nAn article about an organization called a ärtificial revolutionhas been written by someone else. An article about how governments can manipulate people't get rich by pretending to be the ones who want to make America great again. But the statement is not true. It says that the state of affairs must change. To fix the problem, you need to write another letter that tell your friends about the make america Great Againin public. The statement makes america GREAT again protrump! In this blog, we have a poster that proclaims something positive for all of us. However, the statement does not necessarily reflect the views of the nation. People may have different ideas about what makes america fantastic again better than others. This post is not true. Instead, the statement does not necessarily represent merely a promise from the president and a new leader on the continent like it is important to them.
Ground Truth	Fake news

Table 4: Example of rationale generation. The second rationale has been generated with the parameter repetition_penalty set to 1000.

Title	this doctors office has extra wide seats
Image	
Prompt 1	Question: Is this news fake or not? Context: <TEXT> nOptions: <OPTIONS>
Rationale 1	Solution: A news item is a statement that tells you something that is happening at a news organization.nThe news has extra wide seats. The news is not true.nFake news is news that someone is threatening to testify.nThe news has been rewritten to prove that the person is a real person.nThe news has been rewritten to prove that the person is a real person. People often use fake news to find out about something that is happening at the news organization. The news organization has extra wide seats. So, this news is fake..
Prompt 2	Question: Is the statement in this post (A) True or (B) False? <TEXT> Context: Options: <OPTIONS>
Rationale 2	Solution: The statement above is true.nThis doctor's office has extra wide seats. And, the seat width is not correct. Because this post says that the doctors office has extra wide seats, the statements are false! In fact a person would never say that the seats are too narrow. But, the statement is true. It does not state that the doctor's offices have extra wide seats.
Ground Truth	Real news

Table 5: Example of rationale generation. The second rationale has been generated with the parameter repetition_penalty set to 1000.

Title	horses ways of impressing a lady
Image	
Prompt 1	Question: Is this news fake or not? Context: <TEXT> nOptions: <OPTIONS>
Rationale 1	<p>Solution: Facts are words or phrases that use language in a nonliteral or unusual way. They can make writing more expressive.nAnaphora is the sound of a song or a sound. Some people say it is a real thing. But it can also be used to de-emphasize feelings.nFor example, the song I'd better get home before I turn into a pumpkin!was false.nTrue. The sound of the news is not true. It is a fake news story. People often use fake news when they impress someone. Horse's ways of impressing a lady is a real thing..</p>
Prompt 2	Question: Is the statement in this post (A) True or (B) False? <TEXT> Context: Options: <OPTIONS>
Rationale 2	<p>Solution: The statement above is true.nThe horse's ways of impressing a lady are not True.nThis post tell you about horses ways to impress if he or she does not show any emotion at all. This statement is false. Because the statements below are false, they do not express any feelings whatsoever.nTrue! If you ask these questions, the answer is incorrect.nHorses ways Of Impressiving A Lady is an example of how horses can impress adelines. You should be impressed by their looks. But this post says that horses will not impress lords. Instead, try saying that they will not impress a woman.?</p>
Ground Truth	Fake news

Table 6: Example of rationale generation. The second rationale has been generated with the parameter repetition_penalty set to 1000.

Title	obama in letter to america you made me a better man
Image	
Prompt 1	Question: Is this news fake or not? Context: <TEXT> nOptions: <OPTIONS>
Rationale 1	Solution: A news release is a public relations document that tells you something that is true. People often use the words we are united. To identify the differences in the two statements, you need to identify the differences. An interview with the head of Homeland Security was a very positive experience. People often describe the interview as a positive experience. The head of Homeland Security confirmed that the leak was false. But the statement does not state that the leak was fake. The statement tells you that the president has a letter confirming the fact that he has been confirmed. The letter tells you that the president has a letter to america. So, this news is not fake..
Prompt 2	Question: Is the statement in this post (A) True or (B) False? <TEXT> Context: Options: <OPTIONS>
Rationale 2	Solution: The statement above is False. Obama in letter to america you made me better man is false. This post tells true information that obama in letter To America has made me more than one person. But it does not say that obama made me able to do anything. So, the statements are false! In fact no americans have written letters saying that if they were president, then they would be able much better..
Ground Truth	Real news

Table 7: Example of rationale generation. The second rationale has been generated with the parameter repetition_penalty set to 1000.

Prompt	Vision Feature	Use Rationale	Accuracy
Question: Is this news fake or not? Context: <TEXT> Options: <OPTIONS>	No	No	41.8
	DETR	No	41.8
	DETR	RATIONALE-1	44.6
Question: Is the statement in this post (A) True or (B) False? <TEXT> Context: Options: <OPTIONS>	No	No	61.0
	DETR	No	61.2
	DETR	RATIONALE-2	49.5
Question: Is the statement in this post (A) True or (B) False? Context: This post features a piece of text that makes a specific statement or claim about a topic. The statement may be related to politics, science, health, or any other field. The post has been shared widely on social media. <TEXT> Options: <OPTIONS>	No	No	61.1
	DETR	No	61.3
	DETR	RATIONALE-1	49.6
Question: Does the headline accurately reflect the content of this article? Context: (Select option A for True, or option B for False) <TEXT> Options: <OPTIONS>	No	No	60.8
	DETR	No	60.5
	DETR	RATIONALE-1	43
Question: Is the source of this information reliable? Context: (Select option A for True, or option B for False) <TEXT> Options: <OPTIONS>	No	No	60.7
	DETR	No	60.1
	DETR	RATIONALE-1	42
Question: Is the information in this post accurate? Context: (Select option A for True, or option B for False) <TEXT> Options: <OPTIONS>	No	No	60
	DETR	No	60
	DETR	RATIONALE-1	44.9
Question: Is the author's argument based on factual evidence or personal opinion? Context: (Select option A for Factual Evidence, or option B for Personal Opinion) <TEXT> Options: <OPTIONS>	No	No	58.2
	DETR	No	58.2
	DETR	RATIONALE-2	52.1
Question: What is the intended audience for this article? Context: (Select option A for the correct audience, or option B for an incorrect audience) <TEXT> Options: <OPTIONS>	No	No	56.3
	DETR	No	58
	DETR	RATIONALE-2	50.2
Question: What is the author's tone towards the topic? Context: (Select option A for Positive, option B for Negative, or option C for Neutral) <TEXT> Options: <OPTIONS>	No	No	56.0
	DETR	No	58.2
	DETR	RATIONALE-2	49.8

Table 8: Experiments on Fakeddit Dataset with pre-trained mm-cot model

References

- [1] Ashish Vaswani et al. “Attention Is All You Need”. In: (). URL: <https://arxiv.org/abs/1706.03762>.
- [2] Pan Lu et al. “ScienceQA”. In: (). URL: <https://scienceqa.github.io/>.
- [3] Pan Lu et al. “ScienceQA”. In: (). URL: https://lupantech.github.io/papers/neurips22_scienceqa.pdf.
- [4] Zhuseng Zhang et al. “Multimodal Chain-of-Thought Reasoning in Language Models”. In: (). URL: <https://arxiv.org/pdf/2302.00923.pdf>.
- [5] Zhuseng Zhang et al. “Multimodal Chain-of-Thought Reasoning in Language Models”. In: (). URL: <https://github.com/amazon-science/mm-cot>.
- [6] William Yang Wang§ Kai Nakamura Sharon Levy. “Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection”. In: (). URL: <https://arxiv.org/pdf/1911.03854.pdf>.