"Truth is rarely pure and never simple": Fact-checking in Politics

Anonymous ACL submission

Abstract

Veracity detection has emerged as a crucial NLP task over the last decade, as misinformation spreads rapidly in the digital age. Most datasets available in the community, such as LIAR (Wang, 2017) and PolitiFact (Alhindi et al., 2018), either lack complete evidence or have a limited number of instances. We present a new dataset, *Politifact-PLUS*, contains *claim*, *evidence*, *speaker*, *label* and designed for the task of 5-class veracity detection, with particular focus on the political domain. Our analysis examines the efficacy of large language models (LLMs) using prompting approaches, alongside a multi-agent task decomposition framework for veracity detection on our dataset. Notably, we found that the few-shot prompting technique achieved the highest F1 score of **0.7603**, while the task decomposition approach yielded an F1 score of **0.6611**. Our findings highlight the significant confusion among the classes of Mostly True, Half True, and Mostly False. We hope this work inspires the community to develop more robust techniques for veracity detection.

1 Introduction

The proliferation of online information has accelerated the spread of both factual and misleading content, making it increasingly difficult for the public to discern truth from falsehood. In response to this growing challenge, fact-checking platforms like *PolitiFact*[1] have developed systems such as the *Truth-O-Meter*[2], which shows verdict for the claims into varying degrees of accuracy, from **True** to **Pants on fire** and intermediate stages like **Mostly True**, **Half True**, **Mostly False**, and **False**[2]. These labels reflect the complexity of misinformation, where claims often contain elements of truth mixed with misleading or omitted details.

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

[2] https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/

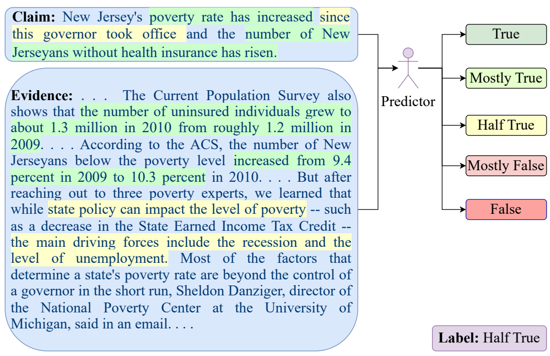

Figure 1: Given a claim and corresponding evidence, the predictor assigns one of the five veracity labels to the claim. Highlighted text in the claim and evidence shows the facts and evidence relationship whether the highlighted content in evidence supports (green) or refutes (yellow) the claim. Since the average evidence length is approximately 760 words, only selective lines are included to maintain relevance and clarity.

Half-truths, in particular, present a unique challenge. They selectively expose the truth and may also contain some false information, exploiting human cognitive biases to manipulate perceptions (Estornell et al., 2020). Unlike outright falsehoods, which are often easier to detect, half-truths thrive on ambiguity. This makes them highly effective in shaping public opinion, particularly in areas like politics, advertisement, and finance, where they are strategically employed to influence decision-making.

Fact-checking is a laborious process that requires significant time and effort. Journalists need to sift through multiple sources to verify claims, assess the credibility of those sources, and draw meaningful comparisons. This process, which can take several hours or even days for professional fact-checkers (Hassan et al., 2015), is often further strained by tight deadlines, especially for internal fact-check procedures (Godler and Reich, 2017). Research indicates that less than half of the pub-

1. Does this claim leave out crucial information considering the evidence?
2. Does the given claim contain false information given the evidence?
3. Is the given claim taking meaning out of context based on the evidence?
4. Does the claim show ambiguity given the evidence?
5. Does the given claim exaggerate based on the evidence?
6. Does the given claim generalize the context based on the evidence?
7. Is the given claim misleading considering the evidence?
8. Does the given claim make a completely ridiculous statement given the evidence?
9. Does the given claim make a false statement given the evidence?

Figure 2: Questions inspired by Truth-o-meter label guideline, politifact.com, described in Appendix, Figure 4.

lished articles undergo verification (Lewis et al., 2008). With the rapid pace of information generation and dissemination, manual fact-checking alone is not scalable, highlighting the need for automation (Guo et al., 2022).

We refer to the work of Frank and Hall (2001) to motivate our approach to fact-checking as a traditional classification task. While fact-checking could be framed as an ordinal classification task, where claims based on evidence are ranked by degrees of truthfulness, this approach faces challenges because accurately capturing the subtle distinctions in a statement's intent and context is not always straightforward. Thus, by simplifying it into a prediction problem, we focus on automating evaluation and measuring performance more effectively.

Figure 1 demonstrates the predictor as a classifier based on evidence that classifies the claim into one of the 5 categories from *True* to *False*, with labels like *Mostly True*, *Half True*, and *Mostly False*. A *Mostly True* claim may omit minor details but remains unambiguous, whereas a *Mostly False* claim significantly distorts the truth or omits some details while keeping small factual elements. *Half true* claim remains in between *Mostly True* and *Mostly False*.

The majority of fact-checking datasets available today contain fewer classes (Thorne et al., 2018; Jiang et al., 2020; Schuster et al., 2021; Hanselowski et al., 2019), typically two or three. This limits the ability to effectively discriminate between varying degrees of the falseness of a claim. More nuanced distinctions are necessary between *True* and *False*, especially in political statements. Existing datasets with four or more classes, such as PUBHEALTH (Kotonya and Toni, 2020) and AnswerFact (Zhang et al., 2020), are from non-political domains, which makes them unsuitable for our focus on political fact-checking. Datasets that do originate from the political domain, such as Liar (Wang, 2017) and PolitiFact (Vlachos and Riedel, 2014), contain either fewer instances to generalize across a five-class classification distribution or doesn't contain complete evidence, which limits their usefulness for multi-class classification tasks.

We propose the Politifact-PLUS dataset, which contains 21,102 instances related to the political domain across five classes of truthfulness. Additionally, Wang (2017) obtained a Cohen's kappa score of 0.82 using data from PolitiFact.com, demonstrating the reliability of the source. This new dataset offers the volume necessary to improve model performance in multi-class political fact-checking.

We conduct various experiments shown in table 3 and table 4 using language models on the Politifact-PLUS dataset for 5-class veracity detection.

Our contributions are:

1. *PolitiFact-PLUS* an extension of Politifact dataset (Misra, 2022), contains extracted fact-checking articles provided by the fact-checkers. We release this extended dataset containing 21,102 instances for the benefit of the research community Section 3.

2. A novel multi-agent framework Section 4.3 boosts the zero-shot F1 score by 10%, which utilizes task decomposition, formulated 9 questions shown in Figure 2 curated from the label description of Politifact's[1] Truth-o-meter[2] verdict label description, for the task of veracity detection.

3. Experiments using LLMs were conducted for the task of 5-class veracity detection on the proposed Politifact-PLUS dataset. We explore approaches including, zero-shot prompting (Kojima et al., 2024), few-shot prompting (Brown et al., 2020), fine-tuning, and 2-stage chain-of-thought reasoning (Kojima et al., 2024). We achieve a 0.7603 F1 score through few-shot learning, showing a considerable performance. Mistral-7B-v0.3 emerged as the top performer Table 3.

2 Related Work

In this section, we review the existing fact-checking datasets listed in Table 1. Following the intuition from Hu et al. (2022), we categorize datasets into two groups: natural (comprising real-world claims) and synthetic (artificially generated). Our focus is on English-language datasets.

Datasets	Type	Domain	#Claim	Meta/Text	#Class
HOVER (Jiang et al., 2020)	Synthetic	Multiple	26,171	Text	2
FEVER (Thorne et al., 2018)	Synthetic	Multiple	185,445	Text	3
VitaminC (Schuster et al., 2021)	Synthetic	Multiple	488,904	Text	3
PunditFact (Rashkin et al., 2017)	Natural	Multiple	4,361	Meta	2/6
Snopes (Hanselowski et al., 2019)	Natural	Multiple	6,422	Text	3
SciFact (Wadden et al., 2020)	Natural	Science	1,409	Text	3
PUBHEALTH (Kotonya and Toni, 2020)	Natural	Health	11,832	Text	4
PolitiFact (Vlachos and Riedel, 2014)	Natural	Politics	106	Text	5
AnswerFact (Zhang et al., 2020)	Natural	Product	60,864	Both	5
LIAR (Wang, 2017)	Natural	Politics	12,836	Meta	6
LIAR-PLUS (Alhindi et al., 2018)	Natural	Politics	12,836	Both	6
Politifact (Misra, 2022)	Natural	Politics	21,152	Meta	6
Politifact-PLUS	Natural	Politics	21,102	Both	5

Table 1: Comparison of different fact-check datasets in English based on their type, domain, number of claims, meta-data/textual-evidence, and number of classes.

2.1 Meta-Based Datasets

Several fact-checking datasets rely solely on claims and their associated metadata. Notably, the LIAR dataset (Wang, 2017) includes metadata such as the claim's speaker, the media source, and a history of the speaker's claims. Similarly, Rashkin et al. (2017) leveraged claims with minimal textual evidence and meta-information. Another early effort by Vlachos and Riedel (2014) resulted in a small dataset that collected claims and meta-information from Channel 4's fact-checking blog[3] and PolitiFact[1]. While these datasets provide related context, they lack evidence to validate claims, limiting their utility for fact-checking tasks.

2.2 Text-Based Datasets

Many text-based datasets focus on Wikipedia as a single source of truth. For example, Schuster et al. (2021) relied solely on Wikipedia, which, while useful, fails to capture misinformation spread beyond what is available on Wikipedia. Datasets such as HOVER (Jiang et al., 2020) and FEVER (Thorne et al., 2018) similarly use only Wikipedia as their knowledge base. Although these datasets provide large-scale examples, they limit their scope to a single source and ignore the varied contexts in which information is interpreted.

To address these limitations, other datasets have been proposed that incorporate evidence from a broader range of real-world sources. These include works by (Hanselowski et al., 2019), (Wadden et al., 2020), (Kotonya and Toni, 2020), and (Vlachos and Riedel, 2014), which provide claims grounded in natural domains beyond politics. However, available political domain datasets provide fewer instances that contain evidence to support or refute the claim, reducing their effectiveness for fact-checking in the political domain.

A significant limitation of LIAR PLUS dataset (Alhindi et al., 2018) is that as evidence, it uses the last 5 sentences of the source article or provides human-written justifications if available. Selected last 5 sentences need not contain complete information to support or refute the claim. This shortcoming limits the dataset's ability to support accurate fact-checking.

In contrast, Misra (2022) introduced a dataset that includes only the claim and metadata from PolitiFact, without incorporating the corresponding retrieved evidence. Building on this, we propose **PolitiFact-PLUS** dataset, which includes not only claims and metadata but also the complete evidence from the PolitiFact website. This enriched dataset provides complete evidence for the veracity detection of the claim.

3 Politifact-PLUS: An Evidence Retrieved Politifact Benchmark Dataset

We address key limitations observed in existing datasets like LIAR (Wang, 2017) which only contains metadata and lacks detailed justification for claims whereas LIAR-PLUS (Alhindi et al., 2018) attempts to extend this by including justifications,

[3] http://blogschannel4com/factcheck/

Label	Train Instances	Test Instances	Validation Instances	Total Instances
True	1,717	612	125	2,454
Mostly True	2,332	824	169	3,325
Half True	2,502	910	179	3,591
Mostly False	2,409	829	184	3,422
False	5,811	2,101	398	8,310
Total	**14,771**	**5,276**	**1,055**	**21,102**

Table 2: Politifact-PLUS Dataset Statistics

but the quality of these justifications is inconsistent. Specifically, LIAR-PLUS either provides human-written justifications or, when unavailable, defaults to extracting the last five lines of the source article. This approach often results in irrelevant or incomplete explanations, which may not accurately capture the reasoning behind the label assignment.

In contrast, our dataset ensures that each claim is accompanied by evidence containing full context about the claim's veracity. The PolitiFact dataset, introduced by Misra (2022), consists of 21,152 fact-checked instances sourced from Politifact.com. Each record contains eight attributes: verdict, statement_originator, statement, statement_date, statement_source, factchecker, factcheck_date, and factcheck_analysis_link. The verdict classifies the truthfulness of a claim into one of six categories: *true*, *mostly true*, *half true*, *mostly false*, *false*, and *pants-fire*. The statements come from 13 different media categories, including speech, television, news, blog, other, social media, advertisement, campaign, meeting, radio, email, testimony, and statement.

To extend this dataset, we create the Politifact-PLUS dataset by extracting articles from the URLs provided in the factcheck_analysis_link attribute. After removing instances without valid URLs, the dataset is reduced to 21,102 instances. We retain four key attributes for our dataset: *label* (verdict), *claim* (statement), *evidence* (the extracted article), *speaker* (statement_originator). Additionally, the *false* and *pants-fire* labels were combined into a single *false* label, as both categories represent completely false information, an example from the dataset can be seen in Figure 1, and example from each class can be seen in Appendix, Figure 5, 6, and 7. The Politifact-PLUS dataset contains *both* **meta/text** and has five truthfulness categories: *true*, *mostly true*, *half true*, *mostly false*, and *false*. The dataset is divided into training, testing, and validation sets with the statistics shown in Table 2.

4 Methodology

In this section, we describe the methods used to evaluate the Politifact-PLUS dataset for 5-class veracity detection. Our approach progresses from small language models finetuning to more resource-intensive techniques, LLMs prompting strategies, and a 2-stage task decomposition framework.

4.1 Fine-tuning Language Models

We first explored fine-tuning language models on our dataset for the 5-class classification task. Using smaller models with parameters $\leq 340M$, adapting them to the Politifact-PLUS dataset.

4.2 Prompting Techniques

Zhao et al. (2024) shows LLMs knowledge can be leveraged using prompt-based techniques. We apply a few prompting approaches to guide the model's predictions:

- **Zero-Shot Prompting:** We used large language models (LLMs) directly without any task-specific training, relying on their pre-existing knowledge to make veracity predictions based solely on the claim and evidence (Kojima et al., 2024).
- **Few-Shot Prompting:** We provided the models with five example instances—one for each class—before asking the model to classify new claims (Brown et al., 2020).
- **2-stage Chain-of-Thought (CoT) Prompting:** In work of Kojima et al. (2024), the task is divided into 2 stages, the first stage takes care of reasoning extraction, by simply adding "Let's think step by step" at the end of the

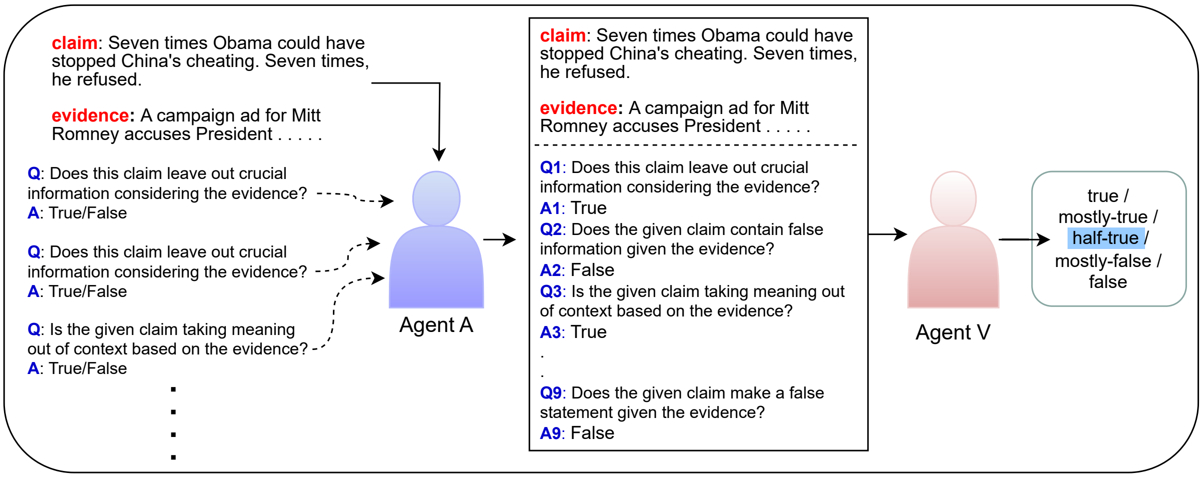

Figure 3: Agent A takes a question, along with the claim and evidence, and generates answers for all 9 questions one by one. Agent V then takes the claim, and evidence, along with all 9 question-answer pairs, and predicts one of the 5 fact-checking labels.

prompt and the second stage takes care of answer extraction by adding "Therefore, among A through E, the answer is" at the end of the response generated after the first stage, to predict the label. Here, labels are encoded as follows A: true, B: mostly-true, C: half-true, D: mostly-false, E: false.

4.3 Multi Agent Task Decomposition Framework

To address the complexity of 5-class classification, we design a multi-agent task decomposition framework inspired by PolitiFact's label guidelines and previous work on decomposing complex claims (Press et al., 2023). The approach breaks down complex queries into simpler sub-queries before answering the main query, which results in an improvement in the success rate of handling compositional tasks. This guides us to design fixed questions that can enhance the prediction accuracy of labels, especially in zero-shot settings. Our framework splits the fact-checking process into two stages, Figure 3.

- **Agent A (Answer Predictor):** In stage 1, Agent A predicts the answers to 9 targeted questions, which are either True or False, as shown in the Figure 2 derived from the PolitiFact[1] truth-o-meter[2] label descriptions. Each question is designed to assess specific aspects of the claim, such as whether it contains false information, is misleading, or leaves out crucial information, along with other aspects.
- **Agent V (Veracity Detector):** In stage 2, using the claim, evidence, questions, and answers generated by Agent A, Agent V predicts the overall truthfulness of the claim. By analyzing the responses to the 9 questions alongside the evidence, Agent V makes an informed decision, classifying the claim into one of the five veracity labels.

5 Experimental Setup

Reynolds and McDonell (2021) demonstrates how prompt structure and wording influence large language models' performance, emphasizing the importance of prompt selection. We utilized the publicly available codebase (Gao et al., 2024) by *EleutherAI*, which offers various evaluation techniques. In our experiments, we employed *log-likelihood-based* evaluation for label prediction and as our dataset is imbalanced, we reported results using the *weighted-f1* score in all cases.

5.1 Models

Our experiments span models ranging from small language models with $\geq 110M$ parameters to large models with $\leq 9B$ parameters. We employ small language models such as BERT (Devlin et al., 2019) (*google-bert/bert-base-uncased*, *google-bert/bert-large-uncased*) and XLNet (Yang et al., 2019) (*xlnet/xlnet-base-cased*, *xlnet/xlnet-large-cased*). These smaller models provide a comparison against the performance of larger models.

For large language models, we utilize state-of-the-art models like Meta's LLaMA (Dubey

	Zero Shot	Few (5) Shot	Zero Shot CoT	Few (5) Shot CoT	QA Task
			Base Models		
Llama-3-8B	0.5471	0.7332	0.3386	0.5465	0.3590
Mistral-7B-v0.3	0.4437	**0.7603**	0.2339	0.6038	0.3381
Gemma-2-9B	0.0674	0.3223	0.1928	0.5887	0.1928
Average	0.3527	**0.6052**	0.2551	0.5796	0.2966
			Instruct Models		
Llama-3-8B-Instruct	0.4425	0.5117	0.2377	0.6122	0.5942
Mistral-7B-Instruct-v0.3	0.4884	0.7330	0.3247	0.6877	**0.6611**
Gemma-2-9B-it	**0.5995**	0.6744	**0.4062**	**0.7065**	0.4229
Average	0.5101	0.6397	0.3228	**0.6688**	0.5594

Table 3: Performance comparison of base and instruct models across various prompting methods—Zero Shot, Few (5) Shot, Zero Shot Chain of Thought (CoT), Few (5) Shot CoT, and QA Task. The results are reported as weighted F1 scores. The 'Average' row shows the average of weighted F1 score for the models under each prompting method.

	True	Mostly True	Half True	Mostly False	False	Overall
			SLMs			
Bert-base-uncased	0.4133	0.3493	0.3348	0.2937	0.7260	0.4953
Bert-large-uncased	0.4349	0.3792	0.3319	0.2854	0.7539	0.5118
xlnet-base-cased	0.4527	0.3977	0.3579	0.3181	0.7436	0.5207
xlnet-large-cased	0.5152	0.4451	0.3950	0.3658	0.7660	**0.5598**

Table 4: Weighted F1 score comparison of small language models (SLMs) including BERT and XLNet (base and large versions) across five veracity classes—True, Mostly True, Half True, Mostly False, and False on the fine-tuning task. The table reports class-specific and overall F1 scores, showcasing the models' performance in the fact-checking task after fine-tuning.

et al., 2024) (*meta-llama/Meta-Llama-3-8B*[4], *meta-llama/Meta-Llama-3-8B-Instruct*[5]), along with Mistral's latest versions at the time of experimentation (Jiang et al., 2023) (*mistralai/Mistral-7B-v0.3*[6], *mistralai/Mistral-7B-Instruct-v0.3*[7]). Additionally, we experiment with models from Google, such as the GEMMA series (Team et al., 2024) (*google/gemma-2-9b*[8], *google/gemma-2-9b-it*[9]).

5.2 Hyperparameter and Prompt Selection

For small language models (SLMs), we experimented with various learning rates: $1e^{-4}$, $5e^{-5}$, $3e^{-5}$, $1e^{-5}$, $5e^{-6}$, $3e^{-6}$, and $1e^{-6}$, using a batch size of 8 till 10 epochs Appendix Table 9. To perform the experiment on the test set, we finalized the *learning rate* = $1e^{-5}$, *batch size* = 8, till *8 epochs*. To perform inference across different approaches such as `zero-shot` (Appendix A.1), `few-shot` (Appendix A.2), and `Chain-of-Thought(CoT)` (Appendix A.3), we conducted experiments on the validation set with various prompts, detailed in Appendix A, results in Appendix table 7, and selected the best-performing prompt for the final results.

6 Results and Error Analysis

Label-Wise Insights: To better understand the overlapping nature of labels that leads to confusion in predictions, we performed a one-vs-all classification experiment. The results, detailed in Appendix Table 8, show the challenges of distinguishing between closely related labels. The task was conducted in two parts: (1) using label-specific

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B
[5] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[6] https://huggingface.co/mistralai/Mistral-7B-v0.3
[7] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3
[8] https://huggingface.co/google/gemma-2-9b
[9] https://huggingface.co/google/gemma-2-9b-it

	True	Mostly True	Half True	Mostly False	False	#Actual
True	566	26	13	0	7	612
Mostly True	200	405	180	33	6	824
Half True	18	61	594	201	36	910
Mostly False	2	8	134	526	159	829
False	5	2	11	137	1946	2101
#Prediction	791	502	932	897	2154	5276

Table 5: Confusion matrix illustrating the number of instances correctly and incorrectly classified across all class labels. Rows represent actual labels, while columns indicate predicted labels. These results correspond to Mistral-7B-v0.3 on a few-shot (5-shot) task.

zero-shot prompts and (2) using a general prompt for all labels (see Appendix B for details).

Interestingly, Mistral emerged as less confused among the classes compared to other models, while GEMMA struggled to perform well in this setting. Across all experiments, the True and False labels performed better than other classes. This trend is likely because it is easier for models to identify claims that are entirely true or entirely false, whereas partially true labels require good reasoning. Label-specific prompt performance is lesser than the general prompt signifies the restrictive nature of providing explicit label descriptions, which may prevent the model from leveraging its generalized, pre-existing knowledge effectively.

The overlapping nature of the labels further adds to the complexity. For instance, a claim missing minor details could be classified as Mostly True or Half True depending on whether the omission creates ambiguity. Determining the correct label requires careful reasoning, as ambiguity and contextual shifts often depend on individual interpretations. Similarly, for claims containing both true and false information, the classification depends on the extent and impact of the false information. A claim is labeled Half True if it becomes ambiguous or misleading but remains majorly true. On the other hand, it is classified as mostly false if the false information predominates, with only minor elements of truth.

Performance Overview:

The performance of SLMs like XLNet-large-cased (*weighted F1*: **0.5598**) and BERT-large-uncased (*weighted F1*: **0.5118**) outperforms their smaller counterparts due to their larger parameter sizes. The False class consistently achieved the highest F1 scores across all models, supported by a higher number of instances (Table 2).

BERT and XLNet use different training strategies: BERT masks 15% of the words randomly and predicts them using surrounding context, while XLNet employs a permutation-based approach to learn contextual dependencies. XLNet's strategy captures better dependencies but both models struggle with complex, overlapping labels like Half True, Mostly False, and Mostly True (Appendix Table 8).

Table 3 compares base and instruct models across prompting techniques (`Zero Shot`, `Few Shot`, `2-stage CoT`, and `QA Task`). The Mistral-7B-v0.3 instruct model in the few-shot setting achieved the best performance, with a weighted F1 score of **0.7603**, a **26.82%** improvement over the zero-shot setting.

We introduce the `QA Task`, a multi-agent task decomposition framework that improves veracity detection by breaking labels descriptions into True/False sub-questions, achieving a **10.27%** performance boost using `Mistral-7B-Instruct-v0.3` over zero-shot setting. This approach, to the best of our knowledge, is novel in veracity detection by simplifying label descriptions through decomposition. Instead of directly asking for predictions, structured questions about the claim-evidence relationship (Figure 2) provide additional context. The answers to these questions, generated by the same model used for prediction, enriched the context and guided the model toward more accurate conclusions.

Interestingly, CoT techniques, particularly in zero-shot settings, underperformed for all models, with an average of **0.2551** for base models and **0.3228** for instruct models, indicating the challenges of logical reasoning without sufficient context or examples. The few-shot CoT task showed a marked relative improvement of **107%**, especially for instruct models, with an average of **0.6688**. Notably, Gemma-2-9B-it outperformed the others in the few-shot CoT task with a score of **0.7065**, showcasing the model's efficiency in handling structured reasoning when primed with examples. CoT technique's reliance on structured reasoning appears to benefit from few-shot learning.

Best Performing Techniques: The Mistral-7B-v0.3 instruct model in the few-shot setting emerged as the best-performing model across all techniques. Its classification report (Table 6) highlights strong performance, particularly for the True and False classes, with F1 scores of 0.8068 and 0.9147, respectively. However, the corresponding confusion matrix (Table 5) reveals that most misclassifications occurred within the Mostly True, Half True, and Mostly False classes, reflecting the inherent ambiguity and overlap between these labels.

For example, 200 instances of the Mostly True label were misclassified as True, and 180 instances as Half True, underscoring the challenge of determining whether omitted information is significant enough to affect the label. Similarly, 201 instances of Half True were misclassified as Mostly False, and 134 instances of Mostly False were misclassified as Half True. This suggests that the model struggles to correctly determine the upper threshold of incorrect information required to classify a claim as Mostly False rather than Half True. This confusion arises because both labels often involve claims containing elements of truth and falsehood, making it difficult for the model to make a clear distinction without deeper reasoning.

As highlighted in the results, the True and False classes consistently achieved higher F1 scores compared to the other labels. This is primarily because completely false information is easier to classify, as it contains no truthful elements, while completely true information is straightforward to identify due to the absence of ambiguity. In contrast, the overlap among classes like Mostly True, Mostly False, and Half True introduces confusion, making it more challenging for the model to distinguish between them.

Conclusion and Future Work

We have expanded the Politifact dataset (Misra, 2022) by incorporating complete evidence from the original sources, resulting in the Politifact-PLUS dataset. Our experiments demonstrate that this additional context significantly improves classification performance. Previous studies, such as Wang (2017) and Alhindi et al. (2018), reported a maximum F1 score of 0.37 across six classes—*true, mostly-true, half-true, barely-true, false,* and *pants-on-fire*—on the LIAR-PLUS dataset, which is a political domain dataset. Our analysis included various prompting techniques and fine-tuning strategies for language models. The best-performing model achieved a **0.7603** weighted F1 score on the Politifact-PLUS dataset. Additionally, by introducing a multi-agent task decomposition framework, we achieved a weighted F1 score of **0.6611**, which represents a **10%** improvement over zero-shot prompting.

In future work, we plan to improve the quality of our generated questions and implement a rule-based system to classify claims based on the answers generated. Furthermore, we intend to explore a range of large language models (LLMs) to further enhance the performance of our fact-checking agents.

Class Label	Precision	Recall	F1-Score
True	0.7155	0.9248	0.8068
Mostly True	0.8068	0.4915	0.6109
Half True	0.6373	0.6527	0.6450
Mostly False	0.5864	0.6345	0.6095
False	0.9034	0.9262	0.9147
Weighted Avg	0.7708	0.7652	0.7603

Table 6: Classification report showing precision, recall, and F1-scores for each class label. The weighted average provides a summary of the overall model performance. These results correspond to Mistral-7B-v0.3 on a few-shot (5-shot) task.

Limitation

This dataset contains instances where the label or label description itself is present so we need to manually clean this dataset as automation can not work because, in the case of removing the words like True or False label, we may lose some relevant information.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings* *of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

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. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. 2020. Deception through half-truths. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10110–10117.

Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. In *Proceedings of the 12th European Conference on Machine Learning*, ECML'01, page 145–156, Berlin, Heidelberg. Springer-Verlag.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Yigal Godler and Zvi Reich. 2017. Journalistic evidence: Cross-verification as a constituent of mediated knowledge. *Journalism*, 18(5):558–574.

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

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Justin Matthew Wren Lewis, Andy Williams, Robert Arthur Franklin, James Thomas, and Nicholas Alexander Mosdell. 2008. The quality and independence of british journalism.

Rishabh Misra. 2022. Politifact fact check dataset.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *XLNet: generalized autoregressive pretraining for language understanding*. Curran Associates Inc., Red Hook, NY, USA.

Wenxuan Zhang, Yang Deng, Jing Ma, and Wai Lam. 2020. AnswerFact: Fact checking in product question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2407–2417, Online. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Prompt Selection

In this section, we present the various prompts explored to identify the most effective one for the 5-class fact-checking task. We also report the weighted F1 scores in table 7 for each prompt evaluated on the validation set, providing insight into the performance differences across the prompt variations.

A.1 Zero Shot Prompts

Base Model Prompts

In this section, we provide the seven prompts used for the base model in the zero-shot setting for the 5-class fact-checking task.

P1
```
Given claim and evidence, predict
if the claim is true, mostly-true,
half-true, mostly-false, or false.
claim: {{claim}}
evidence: {{evidence}}
label:
```

P2
```
Given the evidence, decide if the
given claim is true, mostly-true,
half-true, mostly-false, or false.
claim: {{claim}}
evidence: {{evidence}}
label:
```

P3
```
Given claim and evidence, find if
the claim is true, mostly-true,
half-true, mostly-false, or false.
claim: {{claim}}
evidence: {{evidence}}
label:
```

P4
```
Identify if the claim is true,
mostly-true, half-true, mostly-false,
or false based on the evidence.
claim: {{claim}}
evidence: {{evidence}}
label:
```

P5
```
Given claim and evidence, classify
if the claim is true, mostly-true,
half-true, mostly-false, or false.
claim: {{claim}}
evidence: {{evidence}}
label:
```

P6
```
You need to determine the accuracy of
a claim based on the evidence. Use
one of following 5 labels for the
claim: true, mostly-true, half-true,
mostly-false, or false. Examine the
evidence and choose the most likely
label based on the claim's accuracy
without explaining your reasoning.
claim: {{claim}}
evidence: {{evidence}}
label:
```

P7 Given claim and evidence, you are tasked with evaluating the truthfulness of claims based on the provided evidence. Each claim can be categorized into one of 5 labels: true, mostly-true, half-true, mostly-false, false. Assess the claim given the evidence and classify it appropriately without providing an explanation.
claim: {{claim}}
evidence: {{evidence}}
label:

Mistral Instruct Models Prompts

In this section, we provide the seven prompts used for the Mistral instruct model in the zero-shot setting for the 5-class fact-checking task.

P1 <s>[INST] You are a helpful AI assistant, and you are tasked with evaluating the truthfulness of claims based on the provided evidence. Each claim can be categorized into one of 5 labels: "true", "mostly-true", "half-true", "mostly-false", "false". Assess the claim given the evidence and classify it appropriately without providing an explanation. [/INST] I am excited to work on this classification problem. Can you please provide me with the label description for all 5 labels?
[/INST][Label Descriptions]
true: The claim is accurate and includes all relevant information. There are no omissions or distortions that could mislead the audience.
mostly-true: The claim is accurate, but it might benefit from additional context to provide a complete picture. However, the absence of this context does not alter the claim's accuracy.
half-true: The claim is true in a limited context. However, it omits crucial information that could significantly alter its interpretation, leading to potential misunderstanding or misinterpretation.
mostly-false: The claim contains some elements of truth but distorts or misrepresents critical facts. Important information is omitted, which could lead to a misleading impression despite some truthful elements.
false: The claim is inaccurate and contradicts established facts. The claim has no truth, and it is likely to mislead those who encounter it.
[/INST] Now, can you please provide me with a claim and evidence so that based on the evidence I can classify the claim into one of the 5 labels: "true", "mostly-true", "half-true", "mostly-false", "false".
[/INST]
claim: {{claim}}
evidence: {{evidence}}
label: [/INST]

P2 <s>[INST] Given claim and evidence, you are tasked with evaluating the truthfulness of claims based on the provided evidence. Each claim can be categorized into one of 5 labels: true, mostly-true, half-true, mostly-false, false. Assess the claim given the evidence and classify it appropriately without providing an explanation. [/INST]
Now, can you please provide me with a claim and evidence so that based on the evidence I can classify the claim into one of the 5 labels: true, mostly-true, half-true, mostly-false, false.
</s> [INST]
claim: {{claim}}
evidence: {{evidence}}
label: [/INST]

P3 <s>[INST] You need to judge the truth of a claim based on the evidence given. Use one of these 5 labels for each claim: true, mostly-true, half-true, mostly-false, or false. Review the evidence and classify the claim without explaining your reasoning. [/INST]
Now, can you please provide me with a claim and evidence so that based on the evidence I can classify the

claim into one of the 5 labels: true, mostly-true, half-true, mostly-false, false.
</s> [INST]
claim: {{claim}}
evidence: {{evidence}}
label: [/INST]

P4 <s> Given claim and evidence, you are tasked with evaluating the truthfulness of claims based on the provided evidence. Each claim can be categorized into one of 5 labels: true, mostly-true, half-true, mostly-false, false. Assess the claim given the evidence and classify it appropriately without providing an explanation.
claim: {{claim}}
evidence: {{evidence}}
label:

P5 <s> Given a claim and evidence, you need to decide how accurate a claim is based on the evidence given. Select one of the five labels to classify the claim: true, mostly-true, half-true, mostly-false, or false. Review the evidence, decide how well it supports the claim, and then pick the best label for the truthfulness of the claim.
claim: {{claim}}
evidence: {{evidence}}
label:

P6 <s> You need to determine the accuracy of a claim based on the evidence. Use one of the following 5 labels for the claim: true, mostly-true, half-true, mostly-false, or false. Examine the evidence and choose the most likely label based on the claim's accuracy without explaining your reasoning.
claim: {{claim}}
evidence: {{evidence}}
label:

P7 <s> Given claim and evidence, find if the claim is true, mostly-true, half-true, mostly-false, or false.
claim: {{claim}}
evidence: {{evidence}}
label:

Llama/Gemma Instruct Models Prompts

In this section, we provide the seven prompts used for the LLaMA/Gemma instruct model in the zero-shot setting for the 5-class fact-checking task.

P1 You need to judge the truth of a claim based on the evidence given.
Use one of these 5 labels for each claim: true, mostly-true, half-true, mostly-false, or false.
Review the evidence and classify the claim without explaining your reasoning.
claim: {{claim}}
evidence: {{evidence}}
label:

P2 You need to decide how accurate a claim is based on the evidence given.
Use one of these 5 labels to classify each claim: true, mostly-true, half-true, mostly-false, or false.
Read the evidence, decide how well it supports the claim, and then pick the best label.
claim: {{claim}}
evidence: {{evidence}}
label:

P3 Determine the validity of a claim using the provided evidence.
Select one of the following 5 labels: true, mostly-true, half-true, mostly-false, or false.
Thoroughly review the evidence and accurately categorize the claim without explaining your decision.
claim: {{claim}}
evidence: {{evidence}}
label:

P4 You need to determine the accuracy of a claim based on the evidence.
Use one of the following 5 labels for each claim: true, mostly-true, half-true, mostly-false, or false.
Examine the evidence and pick the most probable label for the claim without explaining your reasoning.
claim: {{claim}}

evidence: {{evidence}}
label:

P5 You need to determine the accuracy of a claim based on the evidence.
Use one of the following 5 labels for each claim: true, mostly-true, half-true, mostly-false, or false.
Examine the evidence and pick the most probable label according to the truthfulness of the claim without explaining your reasoning.
claim: {{claim}}
evidence: {{evidence}}
label:

P6 You need to determine the accuracy of a claim based on the evidence.
Use one of the following 5 labels for the claim: true, mostly-true, half-true, mostly-false, or false.
Examine the evidence and choose the most likely label based on the claim's accuracy without explaining your reasoning.
claim: {{claim}}
evidence: {{evidence}}
label:

P7 Given claim and evidence, you are tasked with evaluating the truthfulness of claims based on the provided evidence.
Each claim can be categorized into one of 5 labels: true, mostly-true, half-true, mostly-false, false.
Assess the claim given the evidence and classify it appropriately without providing an explanation.
claim: {{claim}}
evidence: {{evidence}}
label:

A.2 Few Shot Prompts

Base/Instruct Models Prompts

In this section, we provide the seven prompts used for Base/Instruct models in the few-shot setting for the 5-class fact-checking task.

P1 You need to determine the accuracy of a claim based on the evidence.
Use one of the following 5 labels for each claim: true, mostly-true, half-true, mostly-false, or false.
Examine the evidence and pick the most probable label according to the truthfulness of the claim without explaining your reasoning.
claim: {{claim}}
evidence: {{evidence}}
label:

P2 You need to judge the truth of a claim based on the evidence given.
Use one of these 5 labels for each claim: true, mostly-true, half-true, mostly-false, or false.
Review the evidence and classify the claim without explaining your reasoning.
claim: {{claim}}
evidence: {{evidence}}
label:

P3 Given claim and evidence, you are tasked with evaluating the truthfulness of claims based on the provided evidence.
Each claim can be categorized into one of 5 labels: true, mostly-true, half-true, mostly-false, or false.
Assess the claim given the evidence and classify it appropriately without providing an explanation.
claim: {{claim}}
evidence: {{evidence}}
label:

P4 Given claim and evidence, find if the claim is true, mostly-true, half-true, mostly-false, or false.
claim: {{claim}}
evidence: {{evidence}}
label:

P5 Based on the provided evidence, verify the claim and classify it as true, mostly-true, half-true, mostly-false, or false.
claim: {{claim}}
evidence: {{evidence}}
label:

P6 Based on the provided evidence, judge whether the claim is true, mostly-true, half-true, mostly-false, or false.

claim: {{claim}}
evidence: {{evidence}}
label:

P7 Examine the evidence and classify the claim as true, mostly-true, half-true, mostly-false, or false.
claim: {{claim}}
evidence: {{evidence}}
label:

A.3 CoT Prompts

The Chain of Thought (CoT) method used in this experiment involves a two-stage task. In the first stage, the model generates reasoning by using a trigger phrase: `Let's think step by step`. In the second stage, the model produces the final answer using the phrase: `Therefore, among A through E, the answer is`. The following prompts were used in this setting:

P1 Q: Does the provided evidence support or refute the claim, and how would you categorize the claim based on this evidence?
Answer Choices: (A) True (B) Mostly True (C) Half True (D) Mostly False (E) False.
claim: {{claim}}
evidence: {{evidence}}
A: Let's think step by step.
{{reasoning}}
Therefore, among A through E, the answer is

P2 Q: Does the provided evidence support or refute the claim, and how would you categorize the claim based on this evidence?
Answer Choices: (A) True (B) Mostly True (C) Half True (D) Mostly False (E) False.
Consider category description as follows:
True: The statement is completely accurate, fully supported by the evidence, and does not omit any relevant information that would affect its understanding. There is no ambiguity or need for clarification.
Mostly True: The statement is generally accurate and correct, but it may omit some minor details or require slight clarification. These omissions or clarifications do not affect the core accuracy of the statement but could slightly influence the full understanding of the context.
Half True: The statement is correct in parts but omits crucial information or context that significantly affects understanding. This omission can cause the statement to be misleading or open to multiple interpretations, without necessarily containing false information.
Mostly False: The statement includes a minor element of truth but is largely misleading or inaccurate due to significant omissions, incorrect details, or a distortion of the facts.
False: The statement is entirely inaccurate or false, with misleading or exaggerated claims that deviate significantly from the truth.
claim: {{claim}}
evidence: {{evidence}}
A: Let's think step by step.
{{reasoning}}
Therefore, among A through E, the answer is

P3 Q: Does the provided evidence support or refute the claim, and how would you categorize the claim based on this evidence?
Answer Choices: (A) True (B) Mostly True (C) Half True (D) Mostly False (E) False.
Consider answer choices description as follows:
True: The statement is completely accurate.
Mostly True: The statement is accurate, but it may leave some minor details or require slight clarification.
Half True: The statement is partially correct but omits crucial information or context that affects understanding. This omission can cause the statement to be misleading.
Mostly False: The statement includes

a minor element of truth but is largely misleading or inaccurate due to omissions, incorrect details, or a distortion of the facts.
False: The statement is entirely inaccurate.
claim: {{claim}}
evidence: {{evidence}}
A: Let's think step by step.
{{reasoning}}
Therefore, among A through E, the answer is

P4 Q: Does the provided evidence support or refute the claim, and how would you categorize the claim based on this evidence?
Answer Choices: (A) True (B) Mostly True (C) Half True (D) Mostly False (E) False.
Consider answer choices description as follows:
True: The statement is completely accurate.
Mostly True: The statement is accurate, but it may leave some minor details or require slight clarification.
Half True: The statement is correct but omits crucial information or context that affects understanding. This omission can cause the statement to be misleading.
Mostly False: The statement includes a minor element of truth but is largely misleading or inaccurate due to omissions, incorrect details, or a distortion of the facts.
False: The statement is entirely inaccurate.
claim: {{claim}}
evidence: {{evidence}}
A: Let's think step by step.
{{reasoning}}
Therefore, among A through E, the answer is

P5 Consider claim veracity description as follows:
True: The statement is completely accurate.
Mostly True: The statement is accurate, but it may leave some minor details or require slight clarification.
Half True: The statement is correct but omits crucial information or context that affects understanding. This omission can cause the statement to be misleading.
Mostly False: The statement includes a minor element of truth but is largely misleading or inaccurate due to omissions, incorrect details, or a distortion of the facts.
False: The statement is entirely inaccurate.
Q: Does the provided evidence support or refute the claim, and how would you categorize the claim based on this evidence?
Answer Choices: (A) True (B) Mostly True (C) Half True (D) Mostly False (E) False.
claim: {{claim}}
evidence: {{evidence}}
A: Let's think step by step.
{{reasoning}}
Therefore, among A through E, the answer is

P6 Consider claim veracity description as follows:
True: The statement is completely accurate.
Mostly True: The statement is accurate, but it may leave some minor details or require slight clarification.
Half True: The statement is correct but omits crucial information or context that affects understanding. This omission can cause the statement to be misleading.
Mostly False: The statement includes a minor element of truth but is largely misleading or inaccurate due to omissions, incorrect details, or a distortion of the facts.
False: The statement is completely inaccurate or ridiculous.
Q: Does the provided evidence support or refute the claim, and how would you categorize the claim based on this evidence?

```
evidence?
Answer Choices: (A) True (B) Mostly
True (C) Half True (D) Mostly False
(E) False.
claim: {{claim}}
evidence: {{evidence}}
A: Let's think step by step.
{{reasoning}}
Therefore, among A through E, the
answer is
```

B One vs All

The "One vs All" experiment helps evaluate model performance by isolating each label, allowing us to analyze how well the model distinguishes between one class and the rest. We conduct the experiments results present in table 8 using two types of prompts:

- A label-specific prompt, which incorporates the label descriptions from the PolitiFact Truth-O-Meter[2].
- A general prompt that applies to all labels.

Label specific Prompts

Below are the five specific prompts used for the One vs All experiment:

- **True:** Is the claim fully accurate with no errors or missing information according to the evidence? Respond with `true` or `not true`.
  `claim: claim`
  `evidence: evidence`
  `label:`
- **Mostly True:** Is the claim largely accurate but has minor details missing? Respond with `mostly-true` or `not mostly-true`.
  `claim: claim`
  `evidence: evidence`
  `label:`
- **Half True:** Is the claim partially true but missing crucial information, causing it to be out of context with the evidence? Respond with `half-true` or `not half-true`.
  `claim: claim`
  `evidence: evidence`
  `label:`
- **Mostly False:** Is the claim mostly inaccurate, with the evidence only supporting small parts? Respond with `mostly-false` or `not mostly-false`.
  `claim: claim`
  `evidence: evidence`
  `label:`
- **False:** Given claim and evidence, is the claim entirely inaccurate, with no part of the evidence supporting it? Respond with `false` or `not false`.
  `claim: claim`
  `evidence: evidence`
  `label:`

General Prompt

```
You need to determine the accuracy of
a claim based on the evidence. Use one
of following 2 labels for the claim:
label or not label. Examine the
evidence and choose the most likely
label based on the claim's accuracy
without explaining your reasoning.
claim: {{claim}}
evidence: {{evidence}}
label:
```

Zero Shot							
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**
Base Models							
Mistral-7B-v0.3	0.3213	0.3213	0.3199	0.3396	0.3415	**0.4253**	0.4147
Llama-3-8B	0.29	0.4607	0.4891	0.4678	0.4468	**0.5202**	0.4781
Gemma-2-9b	0.2979	0.3180	0.3264	0.3494	0.3094	0.3473	**0.3769**
Instruct Models							
Mistral-7B-Instruct-v0.3	0.5191	0.5334	0.4060	**0.5428**	0.4832	0.5419	0.5066
Llama-3-8B-Instruct	0.6132	0.4249	0.3550	0.6239	**0.6276**	0.6240	0.4207
Gemma-2-9b-it	0.5183	0.3837	0.4281	0.4041	0.4041	0.3979	**0.5512**
Few(5) Shot							
Base Models							
Mistral-7B-v0.3	0.7690	0.7567	0.7587	**0.7809**	0.7618	0.7778	0.7785
Llama-3-8B	0.6984	0.7123	0.6883	0.7251	0.7304	0.7044	**0.7365**
Gemma-2-9b	0.6566	0.6552	0.6073	0.6914	**0.7127**	**0.7127**	0.6990
Instruct Models							
Mistral-7B-Instruct-v0.3	0.6867	0.6989	0.6856	**0.7360**	0.7215	0.7350	0.7332
Llama-3-8B-Instruct	0.4387	0.4433	0.4908	**0.5505**	0.5235	0.5235	0.5120
Gemma-2-9b-it	0.3700	0.4009	0.3774	0.3625	0.3867	**0.3889**	0.3585
2-stage CoT							
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	
Mistral-7b-v0.3-instruct	**0.5317**	0.4129	0.4339	0.4180	0.4957	0.4604	

Table 7: Weighted F1 Scores for Different Prompts Across Various Models and Experiment Methodologies (Zero-Shot, Few-Shot, and Two-Stage CoT). The scores are reported for multiple prompt configurations for base and instruct models, demonstrating performance variations in prompt selection.

	P	R	F1	P	R	F1	P	R	F1
				Label Specific Prompt					
				Base Models					
	Mistral 7b Base			**Llama 3 8b Base**			**Gemma 2 9b Base**		
True	0.8956	0.1204	**0.0289**	0.8956	0.1223	**0.0327**	0.7192	0.1242	**0.0399**
Mostly true	0.8489	0.5450	*0.5977*	0.8299	0.3242	0.3335	0.8057	0.1716	**0.0687**
Half true	0.8250	0.1905	**0.0924**	0.8592	0.1725	**0.0550**	0.7907	0.3156	0.3186
Mostly false	0.7717	0.3763	**0.4058**	0.8186	0.3109	0.3003	0.7706	0.2104	0.1297
False	0.1423	0.3773	0.2067	0.1423	0.3773	0.2067	0.4723	0.3839	0.2835
				Instruct Models					
	Mistral 7b Instruct			**Llama 3 8b Instruct**			**Gemma 2 9b Instruct**		
True	0.9051	0.6028	0.6697	0.8893	0.6701	0.7285	0.0140	0.1184	**0.0251**
Mostly true	0.8572	0.6910	0.7319	0.8708	0.3318	0.3368	0.8657	0.1697	**0.0633**
Half true	0.8415	0.3583	0.3706	0.8489	0.2815	0.2537	0.8604	0.2076	0.1234
Mostly false	0.7423	0.3592	0.3895	0.8563	0.1829	**0.0691**	0.8145	0.2076	0.1200
False	0.7652	0.3782	**0.2087**	0.7697	0.4085	0.2710	0.7652	0.3782	0.2087
				General Prompt					
				Base Models					
	Mistral 7b Base			**Llama 3 8b Base**			**Gemma 2 9b Base**		
True	0.8993	0.3270	0.3681	0.8895	0.4066	0.4702	0.8956	0.1242	**0.0365**
Mostly true	0.8589	0.3052	0.2998	0.8030	0.5555	0.6124	0.8656	0.1649	**0.0538**
Half true	0.7915	0.2910	**0.2795**	0.8197	0.2550	**0.2137**	0.7590	0.2171	0.1519
Mostly false	0.7293	0.2891	0.2851	0.7688	0.5204	0.5741	0.7338	0.1924	**0.0941**
False	0.7362	0.4218	0.3006	0.7271	0.4227	0.3036	0.5292	0.4009	0.3055
				Instruct Models					
	Mistral 7b Instruct			**Llama 3 8b Instruct**			**Gemma 2 9b Instruct**		
True	0.9199	0.8133	0.8427	0.9086	0.7128	0.7634	0.8965	0.1810	0.1417
Mostly true	0.8739	0.4085	*0.4396*	0.8707	0.4559	0.4977	0.8660	0.1810	**0.0857**
Half true	0.8277	0.6265	0.6728	0.8619	0.3630	0.3734	0.8322	0.2455	0.1953
Mostly false	0.7059	0.1877	**0.0849**	0.8562	0.1801	**0.0634**	0.8565	0.1915	**0.0862**
False	0.7624	0.7555	0.7576	0.7515	0.4834	0.4107	0.6603	0.4284	0.3271

Table 8: Onevsall performance of Base and Instruct Models on validation set using Label-Specific and General Prompts for all, showing Precision (P), Recall (R), and weighted F1 Score (F1) across Different Labels (True, Mostly True, Half True, Mostly False, and False). In comparison to others, Mistral 7b performed better and we got best performance using Mistral 7b Table 3

lr	Bert-base-uncased f1	Bert-large-uncased f1	xlnet-base-cased f1	xlnet-large-cased f1
1e-4	0.1095	0.1095	0.1095	0.1095
5e-5	0.3899	0.1095	0.3349	0.1095
3e-5	0.4051	0.1095	0.4320	0.1095
1e-5	**0.4268**	**0.4408**	**0.4501**	**0.4924**
5e-6	0.4206	0.4340	0.4458	0.4899
3e-6	0.3872	0.4088	0.4509	0.4788
1e-6	0.3086	0.3495	0.4113	0.4580

Table 9: Macro F1 Scores for Different Learning Rates and Models (BERT and XLNet) with a Batch Size of 8, till 10 epochs. We see that at 1e-5 learning rate, we are getting the best result.

Politifact's Truth-O-Meter Guidelines for labels :
True: The statement is accurate and there's nothing significant missing.
Mostly True: The statement is accurate but needs clarification or additional information.
Half True: The statement is partially accurate but leaves out important details or takes things out of context.
Mostly False: The statement contains an element of truth but ignores critical facts that would give a different impression.
False: The statement is not accurate.
Pants on fire: The statement is not accurate and makes a ridiculous claim.

As we clubbed False and Pants on Fire as both contain completely false information only revised definitions are:
True: The statement is accurate and there's nothing significant missing.
Mostly True: The statement is accurate but needs clarification or additional information.
Half True: The statement is partially accurate but leaves out important details or takes things out of context.
Mostly False: The statement contains an element of truth but ignores critical facts that would give a different impression.
False: The statement is not accurate and makes a ridiculous claim.

The intuition behind breaking down into the questions:
True: The statement is accurate and there's nothing significant missing(not misleading, doesn't contain false info, doesn't leave crucial information).
Mostly True: The statement is accurate but needs clarification(may leave crucial information) or additional information.
Half True: The statement is partially accurate(may contain false information) but leaves out important details or takes things out of context making it more generalized.
Mostly False: The statement contains an element of truth(contains false information) but ignores critical facts(leaves crucial information) that would give a different impression(takes things out of context).
False: The statement is not accurate(the claim is false and exaggerated) and makes a ridiculous claim.

Questions arise from label description:
1. Does this claim leave out crucial information considering the evidence?
2. Does the given claim contain false information given the evidence?
3. Is the given claim taking meaning out of context based on the evidence?
4. Does the given claim show ambiguity given the evidence?
5. Does the given claim exaggerate based on the evidence?
6. Does the given claim generalize the context based on the evidence?
7. Is the given claim misleading considering the evidence?
8. Does the given claim make a completely ridiculous statement given the evidence?
9. Does the given claim make a false statement given the evidence?

Figure 4: Given the level descriptions from PolitiFact truth-o-meter guidelines, the key quantities can be extracted from the descriptions. Based on that, the question can be formed.

"id": 16396,
"label": "true",
"speaker": "Florida Democratic Party",
"claim": "(Bill) McCollum also voted for numerous amendments to weaken the legislation, even voting to make trains less accessible to those in wheelchairs.",
"evidence": "Gubernatorial hopeful Bill McCollum has been in politics for much of his adult life, charting a wide map of potential minefields for his enemies to use against him.The Florida Democratic Party, no stranger to partisan warfare, took such aim in a July 26, 2010, press release, using the 20th anniversary of the Americans with Disabilities Act to remind voters of McCollum's initial concerns about the sweeping civil rights legislation.Earlier this year, McCollum said he was proud of his efforts to pass the ADA when he was in Congress, noting there was 'Great Resistance' to the bill. McCollum neglected to state that he was a major player in the 'Great Resistance,' wrote party spokesman Eric Jotkoff in the press release.The press release goes on, He urged then-President George H.W. Bush to reconsider his support of the ADA ... McCollum also voted for numerous amendments to weaken the legislation, even voting to make trains less accessible to those in wheelchairs. McCollum then voted for final passage of the ADA, saying 'politically, it's a very tough vote.' The Americans with Disabilities Act, signed into law July 26, 1990, is considered one of the nation's most important civil rights victories. It requires that disabled Americans be provided reasonable access to employment, transportation, public buildings and communications services. It is widely supported by Republicans and Democrats alike, although some conservatives, libertarians and business groups have long expressed concern that the law puts an undue burden on public servants and employers to accommodate the disabled.The Democrats' claim that McCollum, who represented Florida in Congress for two decades, tried to water down the legislation and didn't want to help people in wheelchairs is an emotional charge. We wondered, is it true?In fact, congressional records show McCollum voted for several amendments to the bill.He supported Amendment 448, which would have specified that the costs required to accommodate the employment of a disabled person not exceed 10 percent of the annual salary or the annualized hourly wage of that job. In debate, McCollum said the amendment, may be the most significant one from the standpoint of mitigating the cost to small business. It failed 187-213 on May 17, 1990. Today, we are going to say that a company manager who earns $40,000 is entitled to a greater accommodation that the mail clerk who receives a salary of $15,000? argued Rep. Dwayne Payne, D-N.J., according to congressional records.On May 22, 1990, the day the legislation passed in the House, records show McCollum voted for Amendment 452, which sought to exempt commuter rail services from the requirement that all new rail cars be readily accessible by persons with disabilities if the commuter rail service made at least one car per train accessible for the disabled. To qualify for the exemption, a rail service would have to add more accessible cars if the demand could not be met by just one car. Proponents of the rail amendment argued that the change offered more specific requirements, albeit different ones, than the original bill, and therefore provided greater protection. For example, the bill required all new purchased or leased buses and rail cars be accessible to the disabled but did not require retrofitting of existing vehicles. The amendment, singled out by the Florida Democratic Party in its press release, failed 110-290. The New York Times wrote at the time, opponents said the effect of the proposed changes would be to segregate people with disabilities.McCollum also voted for Amendment 453 that day, which sought to provide an annual exemption to public transit systems in urban areas with populations of less than 200,000 from the bill's requirement that new vehicles be accessible to people with disabilities, including wheelchair users. To qualify for the exemption, the transit system would have to develop an alternative plan, such as a Dial-a-Ride service. It is the people who need to get from their home to where they want to go, the people who cannot get to the bus stop, are the people who are going to suffer, said sponsor Rep. Bud Shuster, R-Penn. The amendment failed 148-266. A civil right to equal transportation services does not diminish according to a city's population in the latest census, said Rep. Norman Mineta, D-Calif., according to Congressional Quarterly.Finally, McCollum voted for Amendment 454, which sought to keep plaintiffs from suing for monetary damages by limiting the remedies available to discrimination victims to those provided under the Civil Rights Act of 1964, such as injunctive relief, back pay and attorney fees. Supporters argued the disabled should not have greater remedies than those available to women and minorities under the 1964 law. The amendment failed 192-227. You have lesser rights if you have lesser remedies, said Rep. Patricia Schroeder, D-Colo., at the time, according to Congressional Quarterly.On that amendment, McCollum argued at the time: The real problem I have had with the ADA bill altogether, and I am going to vote for this bill, even though a lot of people think that I am out here with a lot of amendments and I am opposed, I am not, because I think the disabled need to have a civil rights bill like this one. I think the problem we have had all along has been costs. It has been a question of how do we mitigate, how do we reduce, costs. It is far more complex under the civil rights legislation for the handicapped than it is for race or sex or any other kind of discrimination. There may be as many as 900, some people say, 900 different disabilities covered by this legislation. There are innumerable different situations in the workplace where the handicapped of different types will intermesh, and those situations will have to be resolved hopefully through processes that are short of litigation. It will be expensive, and even if there is a resolution occasionally by litigation, that will undoubtedly be a very expensive route.McCollum then joined the majority to pass the legislation in a 403-20 vote.So, would the amendments he supported have weakened the legislation? The McCollum campaign did not respond to our questions on this point.But many experts on disability law said the amendments did attempt to undercut the bill.These amendments sought to narrow the rights provided to individuals with disabilities, said Ruth Colker, author of The Disability Pendulum: The First Decade of the Americans with Disabilities Act and a law professor at Ohio State University.The amendments were more pro business and anti-worker, said Paul Steven Miller, a University of Washington law professor and a former commissioner on the U.S. Equal Employment Opportunity Commission, the federal agency that enforces employment discrimination laws. They sort of run counter to what the ADA seeks to accomplish.We also checked whether the law was eventually altered to reflect the amendments supported by McCollum. It wasn't. To this day, Americans with disabilities can sue for monetary damages, there is no fiscal cap on how much an employer may spend to accommodate a disabled employer, and transportation systems still must be accessible.To be sure, the Americans with Disabilities Act does have its critics.After the legislation's passage, the libertarian Cato Institute published a policy analysis by economist Robert O'Quinn, that concluded, the ADA so zealously pursues its mainstreaming goal that individuals, businesses, and governmental bodies must make expensive accommodations to ensure full integration even when less costly, more convenient alternatives, which are preferred by disabled individuals, are available.He continued, The ADA is objectionable on moral as well as economic grounds. In a free society the government should employ its coercive powers only to protect the life, liberty, and property of its citizens from aggression. Any attempt to enforce moral behavior, however noble or desirable, is beyond the proper scope of government.In contrast, some advocates of the bill have complained that it did not go far enough.Our mission, however, is not to judge the value of the acclaimed legislation.The Florida Democratic Party claimed McCollum voted for numerous amendments to weaken the legislation, even voting to make trains less accessible to those in wheelchairs."

Figure 5: An example of a True label instance from the dataset.

"id": 20687,
"label": "mostly-true",
"speaker": "Wisconsin Senate Republicans",
"claim": "Gov. Tony Evers has only gotten one-third of the money meant for COVID relief out the door. He is sitting on $930 million in ARPA funds left unspent. In fact, he still has CARES Act money from two years ago.",
"evidence": "Billions of dollars in federal funding has been flowing into Wisconsin since the early days of the COVID-19 pandemic. The first round of funding arrived in April 2020 courtesy of the Coronavirus Aid, Relief and Economic Security Act (better known as CARES), which dealt nearly $2 billion to the state, to be spent largely at the discretion of Democratic Gov. Tony Evers. The state received another batch in May 2021 as part of the American Rescue Plan Act: About $1.5 billion, with a second payment set to arrive sometime this spring. So far, Evers has directed the money to pandemic-related initiatives such as testing and contact tracing, as well as to broader issues such as infrastructure, tourism recovery and support of small businesses, according to an end-of-year report on the ARPA funds from the state Department of Administration. All of this has given Evers a rare opportunity to dole out money without the approval of the Republican-controlled state Legislature \u2014 and those lawmakers arent happy about it. Theyve pushed unsuccessfully to gain control over how the state should spend the relief money. Now, theyre turning their attention to how fast the governor is getting those dollars out the door. On Feb. 2, 2022, the Senate Republicans made this statement on Twitter: The truth is, Gov. Evers has not acted quickly. He has only gotten one-third of the money meant for COVID relief out the door. He is sitting on $930 million in ARPA funds left unspent. In fact, he still has CARES Act money from two years ago. What is he waiting for? The statement was in response to a tweet from a Democratic state senator, who had praised Evers for acting quickly with the money. There are many things embedded in the Senate Republicans tweet, but were looking here at how much ARPA and CARES Act money has been distributed and how much is still sitting in the states coffers. Lets dig in. When asked to back up the claims, Adam Gibbs, communications director for Senate Majority Leader Devin LeMahieu, sent a Jan. 9, 2022 document from the Legislative Audit Bureau showing that of the nearly $1.5 billion the state had received so far as part of the American Rescue Plan Act, the state had spent about $541 million of it. That would be about a third of that chunk of money \u2014 and leave about $930 million left over. The same document showed Wisconsin has spent about $1.9 billion of the nearly $2 billion in funds from the CARES Act, leaving about $85 million still in Evers hands. On its face, that would make the claim accurate. But theres also a wrinkle. Evers team noted that in addition to the money thats already been spent, there is money that hasnt been spent but has been earmarked for a specific purpose \u2014 in budgetary parlance, this is described as obligated. Any small business owner knows that any accrued expenses should be considered spent, Evers communications director Britt Cudaback wrote in an email. Many pandemic relief programs dont provide funding at the time the award is given, Cudaback said. For example, funds from the states Workforce Innovation grant program are given to grantees periodically as they show progress toward their goals. When looking at how much funding from ARPA was expended or obligated, Cudaback said thats nearly $750 million through Dec. 31, 2021, which doesnt include programs that have been announced since the start of this year \u2014 such as grants for investment in tourism and employee development in the meat-processing industry, among others. Deadlines for allocating and spending the ARPA money also wont approach for years, Cudaback added. Wisconsin must allocate the money by the end of 2024 and spend it by the end of 2026. Wed contend that having nearly $750 million of funds expended or obligated in less than a years time with funds that effectively have a five-year runway meets the definition of acting quickly, she wrote. Similarly, Cudaback said, all $2 billion in CARES Act funding has been allocated, with just about 1% left to be spent. Still, while that provides context on why more money hasnt headed out the door, it doesnt dispute what Senate Republicans claimed. Evers does still have ARPA and CARES Act funding to dole out, even if some of it is earmarked. Senate Republicans claimed that Evers had only gotten a third of COVID relief money out the door, still sitting on about $930 million in ARPA funds, as well as some CARES Act funding. According to the Legislative Audit Bureau, those numbers are right \u2014 but they dont take into account the fact that some of those funds are already set aside, or obligated, to specific purposes."

"id": 18350,
"label": "half-true",
"speaker": "Barack Obama",
"claim": "We've brought trade cases against China at nearly twice the rate as the last administration.",
"evidence": "President Barack Obama touts his administrations record holding trade partners accountable by drawing a contrast with President George W. Bush over China.We've brought trade cases against China at nearly twice the rate as the last administration, he said in an April 13, 2012, speech in Tampa, Fla., before a trip to Colombia.Heres how he set it up: Now, one of the ways that we've helped American business sell their products around the world is by calling out our competitors, making sure they're playing by the same rules. For example, we've brought trade cases against China at nearly twice the rate as the last administration. We just brought a new case last month. And we've set up a trade enforcement unit that's designed to investigate any questionable trade practices taking place anywhere in the world. Its a claim hes made before, published in the Los Angeles Times and the New York Times.We wondered, is it true?The honeymoonWe asked the White House for support for the presidents claim.Obama referred to cases brought against China before the World Trade Organization, said spokesman Matt Lehrich. The WTO is a group of more than 150 governments that sets and enforces international trade rules. Since the United States and China are both members, its a pivotal place they can go to settle disputes with one another.There are other important types of trade cases, such as anti-dumping cases brought before the U.S. International Trade Commission, but those are brought by private industry, said Peg OLaughlin, public affairs officer for the ITC.Other kinds of enforcement cases include those brought under Section 301 or 201 of U.S. trade laws. They're rare now because, under WTO rules, the United States isn't supposed to regularly turn to that sort of unilateral action, said Paul Blustein, a trade expert with the Brookings Institution.So while there are a wide range of trade measures available, the experts we consulted said focusing on just WTO cases seemed reasonable.The Obama administration has brought six cases against China before the WTO in less than one term, while the Bush administration brought seven cases over two terms \u2014 thus, the claim at nearly twice the rate.But theres a distinction between the two presidencies, what well call Chinas honeymoon.China joined the WTO in 2001, after Bush took office. At that point, member countries essentially gave China a grace period to follow the new rules.Business was just rushing into China those were good days, said Gary Clyde Hufbauer, a senior fellow for the Peterson Institute for International Economics who writes about U.S.-China trade and worked in the Carter and Ford administrations. Nobody was wanting to bring cases in particular. (China) probably got more of a grace period than would normally be expected because of the business boom.The United States became the first country to file a trade case, over trade barriers against integrated circuits, in March 2004.Jerry Jasinowski, president of the National Association of Manufacturers, said in 2004 that China had needed time to adjust its tax and regulatory policies to comply with WTO standards, but that after two years the honeymoon was over, according to Congressional Quarterly and other news organizations.The Obama administration, on the other hand, had no such delay \u2014 plus it could take advantage of work started under Bush to file a first case within six months of taking office.Thats not to say Democrats back in 2004 werent arguing the Bush administration could have acted sooner.One case brought as a political talking point does not make up for the administration's failure to develop a China trade policy over the past three years, Rep. Sander Levin of Michigan, the top-ranking Democrat on the Ways and Means Trade Subcommittee, was quoted as saying in Congressional Quarterly in 2004. This is an open-and-shut case that the administration should have addressed years ago.A Bush-era deputy U.S. trade representative says Bush realistically had five years to bring cases against China \u2014 not seven.The first year of China's membership was eaten up giving them a chance to prove compliance, said John Veroneau, who also worked in the Defense Department under President Bill Clinton. The second year was eaten up jawboning about problems and preparing the facts and analysis to be able to bring a WTO case. At the beginning of the third year, we brought the first case.That changes the math, putting Bushs rate of cases much closer to Obamas.Rather than nearly two cases a year for Obama vs. about one for Bush, the comparison would be nearly two cases a year for Obama vs. about one and a half for Bush.Driving factorsThe trade policy of the president isnt necessarily the largest factor driving the rate of trade cases, said Hufbauer, the expert with the Peterson Institute. Other considerations out of Obamas control (and Bushs) held greater sway, he said."

Figure 6: Instances of a Half True and Mostly True label from the dataset.

"id": 182,
"label": "mostly-false",
"speaker": "Mark Pocan",
"claim": "Says withdrawing troops from Afghanistan could save the U.S. $50 billion.",
"evidence": "After nearly 20 years of conflict in Afghanistan, the United States has pledged to withdraw its forces from the country. President Joe Biden announced a plan to have all U.S. troops out of the country by Sept. 11, 2021, but an accelerated pace of withdrawal could have troops completely out of the country by mid-July, according to a May 25, 2021 report from the New York Times. As Americas longest war draws to a close, arguments have sprung up over what to do with the money the country is currently spending on the conflict -- and how much will actually be saved by pulling troops out of the unstable region. U.S. Rep. Mark Pocan, D-Madison, claimed that withdrawing troops from Afghanistan could save the country $50 billion a year in a May 21, 2021 tweet -- money he argues could be cut from the Pentagon budget and put towards something else, such as ending homelessness. For the purpose of this fact check, were going to focus on the first part of his claim. Can bringing home the troops in Afghanistan really save the country $50 billion? In short, not as a practical matter. When we reached out to Pocans office seeking backup, communications director Usamah Andrabi said that the $50 billion had been widely reported, and shared a link to a report by The Balance, a nonpartisan financial advice and news site, based in New York City. The war started off in 2001, with 9,700 people on the ground in Afghanistan, at a cost of $23 billion, according to The Balance. That number has grown since then, hitting $107 billion in spending in 2011, with more than 94,000 people on the ground. Since then, yearly spending has dropped as the number of troops stationed in the country has declined. In 2018, that number dropped to $52 billion in spending, and remained the same for 2019, according to an estimate by The Balance. Spending for 2020 was not yet available. But even though the U.S. is currently spending about $50 billion a year on the war, that doesnt mean that pulling troops out will amount to that same figure in savings. Jonathan Bydlak, director of the Governance Program for the R Street Institute, a nonpartisan policy research organization, said estimating cost savings from shifts in ground troops and other foreign policy decisions isnt straightforward. There are three things that would need to be considered to reach an estimation of savings, he said: The direct costs of engagement. Changes in the base Department of Defense budget because of reduced engagement. Ongoing/long-term costs, primarily veterans medical/disability benefits and interest. Bydlak estimated the U.S. could see about $4 billion to $6 billion in direct savings, about $1 billion to $2 billion in base budget savings and about $28 billion to $42 billion in long-term costs. That puts total savings somewhere between $33 billion and $50 billion a year. So, Pocans claim is on the very high end of that range. But that savings could shrink if Biden opts to only withdraw a portion of the troops currently on the ground, leaving a small residual force. In that case, savings would only be about $7 billion to $10 billion. There are also other costs that could crop up, too, Bydlak said: If the U.S. decides to provide more aid to Afghanistan, to help encourage stability; if more money is spent by the Department of Homeland Security in the wake of withdrawal; or if the domestic cost of housing troops is greater than the cost of stationing them in Afghanistan, due to a higher cost of living. Others worry that any savings for the U.S. could be eaten up -- at least in the short term -- by the cost of pulling troops and supplies out of Afghanistan. Mackenzie Eaglen, a resident fellow at the American Enterprise Institute, a right-leaning public policy think tank, said in an April 26, 2021 report that spending in Afghanistan will still remain high without boots on the ground due to ongoing investments in counterterrorism and salaries and other expenses for the 3,000 members of the Afghan National Security Forces. Among other potential costs Eaglen included in her report: Breaking contracts with private entities for property, buildings and equipment and bringing home the equipment the U.S. brought with its troops. It will require more forces than are in the country now, the article said. Bringing troops home isnt an end to the mission that started in 2001 in Afghanistan, its a mission change, Eaglen wrote. If Congress is expecting a windfall of savings to result from the Afghanistan withdrawal, it is likely to be disappointed, Eaglen wrote. Threats will still need to be managed -- just from slightly farther away. In the meantime, it will discover that leaving is hard, dangerous, and expensive. Pocan claimed in a tweet that the country could save $50 billion a year by pulling troops out of Afghanistan. The U.S. could save up to $50 billion, or as little as $7 billion on withdrawing troops, according to one expert. But that just covers one side of the ledger."

"id": 5732,
"label": "false",
"speaker": "Barack Obama",
"claim": "The most realistic estimates for jobs created by Keystone XL are maybe 2,000 jobs during the construction of the pipeline.",
"evidence": "The Keystone XL pipeline that would carry oil from Alberta, Canada, to refineries on the Texas Gulf coast presents President Barack Obama with no easy choice. While officially, the final decision to block or approve it is in the hands of the State Department, politically, the plan pits two key Democratic constituencies against each other, environmentalists and organized labor. For the first group, extracting petroleum from Canadian tar sands is a climate change disaster. For the unions, the project means jobs. Fresh off a speech that underscored the need to restore Americas middle class, Obama talked about the steps that lie ahead for the 875-mile link between the Canadian border and a distribution hub in Nebraska. The central question, he said in an interview with the New York Times, is whether this would significantly contribute to carbon in our atmosphere. As for jobs, the president went out of his way to downplay them. In the big picture, they were but a blip, as the president put it. Republicans have said that this would be a big jobs generator. There is no evidence that thats true, Obama said. Any reporter who is looking at the facts would take the time to confirm that the most realistic estimates are this might create maybe 2,000 jobs during the construction of the pipeline -- which might take a year or two -- and then after that were talking about somewhere between 50 and 100 jobs in a economy of 150 million working people. Theres been a running battle over jobs and the Keystone XL. Weve checked claims that it would employ as many as 20,000 workers. To be clear, there are all sorts of complications when it comes to predicting how many jobs a complex, two-year project will generate. There are the direct construction jobs; theres indirect employment at companies that provide the materials and services related to the work; and then theres the really indirect effect that comes when money is pumped into an economy and people buy food and pay rent and so on. But out of all the numbers bruited about, the presidents seemed particularly low. We asked the White House for evidence to support the claim. All they offered was a statement from spokesman Josh Earnest during a press briefing. There are a range of estimates out there about the economic impact of the pipeline, Earnest said. What the president is interested in doing is draining the politics out of this debate and evaluating this project on the merits. During the New York Times interview, the president invited reporters to use the most realistic estimates. So we went to the State Departments lengthy environmental impact statement on the project that came out in March. In that report, the lowest estimate for jobs directly tied to construction was 3,900 jobs a year. That number came after analysts wrestled with the stop-and-start nature of construction work and converted the jobs to a yearly estimate. Approximately 10,000 construction workers engaged for 4-to 8-month seasonal construction periods (approximately 5,000 to 6,000 per construction period) would be required to complete the proposed project. When expressed as average annual employment, this equates to approximately 3,900 jobs. The analysis noted that 90 percent of those jobs would come from a unique national labor force that is highly specialized in pipeline construction techniques. It also confirmed that there would be few long-term jobs, something on the order of 35. The largest jobs number in the State Department report is an annual average of 42,100, but that includes part-time jobs and folds in the ripple effects as spending moves through the economy, measured over two years. The further out from the immediate project the analysis moves, the less certain the results. The report said these jobs would amount to 0.02 percent of total American employment, adding some weight to the presidents characterization of the impact on the overall jobs picture. The North American Building Trades Union said it was disappointed with Obamas words and pressed him to let the pipeline move forward. So that workers and their families can share in the economic recovery he is touting, said union president Sean McGarvey. The president should look to his own State Departments findings that there will be meaningful job creation. We looked at the website of the Sierra Club, one of the leading environmental groups opposed to the pipeline, and they used the State Departments 3,900 annual number. The only place we found anything close to the presidents figure was at the Cornell School of Industrial and Labor Relations Global Labor Institute. Assistant director Lara Skinner co-wrote a report highly critical of the pipeline. Skinner argued that the 3,900 covered employment for two years and that it should be divided in half. That's where the 2,000 job figure comes from, Skinner said. "

Figure 7: Instances of a Mostly False and False label from the dataset.