"Politi-Fact-Only": A Political Domain Benchmark Dataset

Anonymous ACL submission

Abstract

The rapid proliferation of online information has made it increasingly challenging to differentiate factual content from misinformation. Traditional fact-checking methods, which require extensive manual effort, are not scalable given the volume of misinformation spreading online. Automated fact-checking has emerged as a promising solution, leveraging machine learning models trained on datasets derived from fact-checking websites (Wang, 2017; Augenstein et al., 2019; Gupta and Srikumar, 2021a). However, many of these datasets include post-analysis commentary from annotators, which may introduce bias and provide implicit cues that aid model performance. To address this limitation, we introduce Politi-Fact-Only, a benchmark dataset comprising 1,482 instances curated from PolitiFact.com, where we remove post-analysis and retain only factual evidence Fig 1. This ensures that models must rely solely on factual reasoning rather than verdict-related information. Our experiments demonstrate that state-of-the-art fact-checking models, including large language models (LLMs), struggle to accurately classify claims when deprived of post-claim analysis, highlighting their reliance on implicit cues rather than pure factual reasoning.

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], 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**. These labels reflect the complexity of misinformation, where claims often contain elements of truth mixed with misleading or omitted details, whereas half-truths present unique challenges. They selectively expose the truth, 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-

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

claim: A photo shows a crash in Eminence, Indiana.
evidence: ~~With wintry weather striking many regions of the United States, photos are emerging from across the country showing snow plows, icy roads, and big drifts of snow. But one old picture is being mischaracterized as showing the aftermath of a crash in Eminence, Indiana.~~ This is currently at state road 42 and state road 142 in downtown Eminence, Indiana, a Feb. 2 post says, alongside a photo of a treacherous-looking pileup of cars and trucks. Black ice! Drive safe folks! Right by the Citizens Bank and Dairyland. Prayers to all involved. But this photo was taken last year, and about 900 miles southwest of Eminence. ~~This post was flagged as part of Facebook's efforts to combat false news and misinformation on its News Feed. (Read more about our partnership with Facebook.)~~ Photographer Lawrence Jenkins took it on Feb. 11, 2021, in Fort Worth, Texas, where 133 vehicles crashed after freezing rain coated the roads there, sending dozens of people to the hospital and leaving at least six dead, the Dallas Morning News reported at the time. North Texas and Central Indiana are both experiencing wintry weather, ~~but this photo doesn't show it.~~
label: false

source: social_media
speaker: Facebook posts
claim_data: 19/11/2018
factchecker: Jill Terreri Ramos
fact_check_data: 7/12/2018
factcheck_analysis_link: https://www.politifact.com/factchecks/....

Figure 1: An example from the dataset where sentences with strike-through lines represent information added only after the claim was verified. These lines were manually deleted to ensure the evidence contained only factual details sufficient for fact-checking the claim. Also, we show the meta-data available for an instance.

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 published 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).

Several studies have explored automated fact-checking, including works by (Wang, 2017), (Augenstein et al., 2019), and (Gupta and Srikumar, 2021a), which have contributed valuable datasets. While these datasets contain real-world claims, they are primarily derived from fact-checking websites. The articles on such platforms often present a post-analysis of claims, incorporating assessments from annotators based on factual evidence. However, this does not fully reflect real-world fact-checking scenarios. To address this limitation, some researchers, such as (Yang et al., 2022) and (Khan et al., 2022), have focused on utilizing premise articles or sources that were published before the claim itself. This approach brings the fact-checking process closer to real-world settings. However, while relevant information is extracted from these documents based on the claim, there is no guarantee that the retrieved content is sufficient for verification. *To bridge this gap*, we propose a benchmark political domain test set Politifact-Facts-Only Section 4, a subset of Misra (2022). In this dataset, we manually remove the post-analysis provided by annotators and retain only the factual information. This ensures a more realistic evaluation of automated fact-checking models.

Figure 1 illustrates an example from the dataset, where annotators manually reviewed the instances and removed sentences containing post-claim analysis. In real-world scenarios, fact-checking relies solely on factual information, requiring reasoning based on these facts without the aid of post-publication commentary. While previous approaches (Khan et al., 2022) have attempted to address this by using review or premise articles to avoid post-analysis content, this raises a critical concern about the effectiveness of an abstract summary of evidence extracted for accurate fact-checking of claims.

Our contributions are:

1. We introduce *Politi-Fact-Only*, a benchmark dataset comprising 1,482 instances curated from PolitiFact.com[2]. As detailed in Table 1, the dataset has been manually filtered to remove post-claim analyses and verdict-related information originally present in the articles (Section 4). This ensures that the evidence consists solely of factual content, eliminating potential annotator bias introduced by verdict cues and improving the reliability of fact-checking models. Initially, we collected 1,500 instances. However, to maintain accuracy and credibility, we removed 18 instances due to reasons outlined in Section 5.

2. Through experiments we show the performance of *Politi-Fact-Only* along with other various datasets, Table 3 and 2. We observe that on our test set models are struggling to reason about the facts to support or refute the claim. Large language models (LLMs) struggle to reason effectively when limited to fact-only evidence from the *Politi-Fact-Only* dataset. In contrast, LLMs perform comparatively better on other datasets in zero-shot settings, highlighting their reliance on implicit cues and verdict-related information rather than pure factual reasoning.

2 Problem Statement

Input: A **claim** $C = \{c_1, c_2, \ldots, c_n\}$ and its corresponding **evidence** $E = \{e_1, e_2, \ldots, e_m\}$, where c_i and e_j represent individual tokens in the claim and evidence, respectively.

Output: A **verdict label** L, where L belongs to {True, Mostly True, Half True, Mostly False, False}

The goal is to classify the claim C based on the factual content of E, determining its degree of truthfulness.

3 Related Work

Existing fact-checking datasets can be broadly categorized into meta-based and text-based datasets. Meta-based datasets, such as LIAR (Wang, 2017) and (Rashkin et al., 2017), primarily include claims with metadata like speaker identity and historical records but lack supporting textual evidence, limiting their utility for verification. Similarly, Vlachos and Riedel (2014) compiled a small dataset of political claims, but without explicit evidence, restricting its effectiveness in real-world fact-checking.

Text-based datasets offer stronger evidence-grounded verification, with FEVER (Thorne et al., 2018), HOVER (Jiang et al., 2020), relying solely on Wikipedia as a knowledge base. While valuable, these datasets fail to capture misinformation from diverse sources beyond Wikipedia. Some datasets, such as Multifc (Augenstein et al., 2019) and X-fact (Gupta and Srikumar, 2021b), incorporate evidence from broader domains. LIAR-PLUS (Alhindi et al., 2018) attempted to provide evidence by extracting the last five sentences from source articles, but this often resulted in incomplete or irrelevant context. L++ (Russo et al., 2023) and ru22fact (Zeng et al., 2024) make use of fact-checking website for the dataset creation. Khan et al. (2022) and (Yang et al., 2022) worked on more real-world situations by fetching content from the article that exists before the claim was published. To address these limitations, Politi-Fact-Only builds upon PolitiFact (Misra, 2022) by filtering out post-claim analysis and retaining only factual evidence. This ensures a more realistic test set for evaluating models on factual reasoning without relying on annotator cues.

Label	**Count**	**Token$_\mu$**	**Sent$_\mu$**	**BPE$_\mu$**
True	296	596.05	26.05	755.94
Mostly True	298	682.90	30.88	865.19
Half True	293	756.69	33.61	954.85
Mostly False	300	780.14	34.88	978.39
False	295	559.09	27.26	705.01
Total	1482	675.18	30.55	852.13

Table 1: Statistics for *Politi-fact-only* dataset. Token$_\mu$, Sent$_\mu$, and BPE$_\mu$ represent the average number of standard tokens, sentences, and BPE tokens per evidence, respectively.

4 Politi-fact-only: A Fact Only Benchmark Dataset

Our dataset, *Politi-Fact-Only*, ensures that each claim is accompanied by evidence containing facts related to the claim, supporting its veracity. We randomly selected 1,500 instances from the PolitiFact dataset (Misra, 2022), sourced from Politifact.com. Section 5 details our manual filtration and annotation process. We removed instances where the predictions did not match, resulting in a final dataset of 1,482 instances, as shown in Table 1. Upon comparing Table 1 with Table 4 in the Appendix, we observe a decrease of approximately 15% in the average after filtration. This suggests that around 15% of the content in the article consisted of commentary from the annotators.

Each record in *Politi-Fact-Only* contains nine attributes. We retain the following key attributes: *label*, *claim*, *evidence*, *speaker*, *factcheck_analysis_link*, *factcheck_date*, *factchecker*, *claim_date*, and *claim_source*. Additionally, the *false* and *pants-fire* labels were merged into a single *false* label, as both categories represent completely false information. Examples from each class can be found in the Appendix, Figures 2 - 6.

We retain the claim's publish date and the fact-check date for relevant use cases, such as extracting evidence articles using the Google API. By using the claim verbatim, as done by (Yang et al., 2022), we can filter out articles published after the claim's publish date and fact-check date, depending on the research needs. The speaker of a claim is crucial, as it provides insights into their reliability. If a speaker frequently makes false statements, it indicates the speaker's lack of credibility. The same applies to the source, whose history of publications can reflect its credibility.

An example is shown in Figure 1, where strikethrough lines indicate content present solely due to annotator commentary. Our goal is to make this dataset as representative as possible of real-world scenarios, where only factual information is available. To achieve this, we must extract factual content from the web. However, extracting relevant information up to a predefined limit (e.g., k) does not guarantee that the summary will be sufficient to assess the veracity of the claim.

5 Annotation and Filtration Process

We hired three annotators, all proficient in English, to clean 500 instances each. The annotators were instructed to annotate the instances while also applying a filtration process. The evidence for each instance was obtained by scraping fact-checking websites. These articles included post-claim analyses written by fact-checkers from sources such as PolitiFact, who provide ratings ranging from "True" to "Pants on Fire."

As shown in Appendix B, we established specific guidelines for the filtration process and provided some filtered instances as reference. By adding two new fields for each instance: **leaked** and **annotator_prediction** we asked annotators to fill this accordingly. We removed sentences from the evidence if they were directly related to the verdict,

Models	MultiFC	Liar Plus	RU22fact	L++	Politi-Fact-Only	Unfiltered
Mistral	0.1419/**0.2952**	0.2060/**0.2822**	0.3239/**0.6554**	0.2874/**0.3607**	0.2667/**0.3475**	0.3725/**0.4575**
LLaMA	0.1579/**0.2532**	0.1853/**0.2580**	0.2933/**0.6462**	0.2851/**0.3553**	0.2115/**0.2861**	0.4811/**0.5142**
Gemma	0.1586/**0.2540**	0.0708/**0.1723**	0.2710/**0.6427**	0.1845/**0.3107**	0.2057/**0.2821**	0.5904/**0.6005**

Table 2: Performance comparison of models ranging from 7B to 9B parameters using Zero-Shot prompting (Kojima et al., 2024)across various fact-checking datasets. The results are reported in macro F1/micro F1-score. The "Unfiltered" dataset represents the unfiltered version of *PolitiFact-Fact-Only*. We use models like Meta's LLaMA (Dubey et al., 2024) (*meta-llama/Meta-Llama-3.1-8B*), Mistral's version 3 (Jiang et al., 2023) (*mistralai/Mistral-7B-v0.3*)and, model from Google, such as the GEMMA series (Team et al., 2024) (*google/gemma-2-9b*) from huggingface.

Dataset - Model	Respective Dataset	PolitiFact-Fact-Only
LIAR-PLUS - SVM (Alhindi et al., 2018)	0.25	0.27
LIAR - CNN (Wang, 2017)	0.27	0.28
LIAR-PLUS - LR (Alhindi et al., 2018)	0.37	0.27
AVERITEC - BERT-large (Schlichtkrull et al., 2023)	0.49	0.29

Table 3: Comparison of *Politi-Fact-Only* with other fact-checking datasets. All results are reported in Macro F1-score. The values for the respective datasets are sourced from the original authors' reported results.

such as statements like “This post was flagged as part of...”. Similarly, we eliminated annotator commentary, such as “But this photo doesn't show it”. These logical or inference-based cues can make reasoning easier; however, in real-world scenarios, only factual information is available, making it significantly harder to assess the claim's veracity based solely on the provided evidence.

Consequently, we removed 15 instances where the annotator's prediction did not match the expected outcome. Additionally, we excluded three more instances because, after the cleaning process, the reduced context no longer provided enough information to support or refute the claim. By publishing this dataset, we aim to provide a resource that can be used for both explanation generation and fact-checking classification tasks.

6 Experimental and Results Analysis

We conducted two experiments: one utilizing existing fact-checking datasets sourced from PolitiFact.com and another leveraging large language models, including *meta-llama/Meta-Llama-3.1-8B* (Dubey et al., 2024), *mistralai/Mistral-7B-v0.3* (Jiang et al., 2023), and *google/gemma-2-9b* (Team et al., 2024). As shown in Table 3, our dataset outperformed LIAR-PLUS when evaluated using an SVM classifier. Alhindi et al. (2018) primarily extracted the "Our Ruling" or "Our Rating" section as evidence when available; otherwise, they relied on the last five lines of the article. Yang et al. (2022) further refined this approach by retaining only instances containing the "Our Ruling" or "Our Rating" section, using it as gold-standard evidence for comparison with their generated explanations. We also tried different prompts for different models and we analyze that a single keyword can effect LLM's output Appendix A.

Conclusion and Future Work

We introduced *Politi-Fact-Only*, a benchmark dataset for evaluating fact-checking models using only factual evidence, without post-claim analysis. Our experiments show that models struggle significantly when deprived of annotator cues, resulting in a notable performance drop compared to unfiltered datasets. This highlights their reliance on implicit signals rather than pure factual reasoning. While large language models (LLMs) perform well on traditional datasets, their difficulty in classifying claims accurately in our test set suggests dependence on verdict-related information. This underscores the challenge of building robust fact-checking systems that operate without pre-annotated guidance.

For future work, we propose using Politi-Fact-Only as a test set to evaluate retrieved or summarized information from web articles and documents, assessing whether automated methods preserve enough factual content for verification. Additionally, we aim to enhance fact-checking models by incorporating reasoning-driven approaches that rely solely on factual evidence. Addressing these challenges will contribute to the development of more transparent, and effective fact-checking systems for real-world misinformation detection.

Limitation

This dataset is collected from a fact-checking website. While we have attempted to remove most annotator cues, some sentences could not be eliminated without compromising the context necessary to support or refute the claim.

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.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. 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.

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.

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.

Ashim Gupta and Vivek Srikumar. 2021a. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Ashim Gupta and Vivek Srikumar. 2021b. X-FACT: A New Benchmark Dataset for Multilingual Fact Checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. 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.

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.

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.

Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. WatClaimCheck: A new dataset for claim entailment and inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304, Dublin, Ireland. 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.

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.

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.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the Generation of Fact Checking Explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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.

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.

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.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Chang Yi. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 2608–2621.

Yirong Zeng, Xiao Ding, Yi Zhao, Xiangyu Li, Jie Zhang, Chao Yao, Ting Liu, and Bing Qin. 2024. RU22Fact: Optimizing evidence for multilingual explainable fact-checking on Russia-Ukraine conflict. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14215–14226, Torino, Italia. ELRA and ICCL.

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 5 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:
```

Label	Count	Token_μ	Sent_μ	BPE_μ
True	296	666.20	29.40	847.27
Mostly True	298	804.78	36.51	1022.50
Half True	293	898.12	39.95	1135.38
Mostly False	300	927.16	41.40	1165.60
False	295	679.12	33.05	862.42
Total	1482	795.31	36.07	1006.92

Table 4: Statistics for Unfiltered version of Politi-fact-only dataset. Token_μ, Sent_μ, and BPE_μ represent the average number of standard tokens, sentences, and BPE tokens per entry, respectively.

B Guidelines

I met with the annotators regularly over a span of two months. During this time, we employed three annotators who were proficient in English and compensated by our lab. The Politifact-only dataset is a fact-checking dataset scraped from politifact.com, focusing on the political domain. It consists of 1,500 instances, each containing a political claim along with corresponding evidence.Based on the evidence, the claim's truth value is categorized in one of the following categories: true, mostly true, half true, mostly false, false, pants on fire. I have

	P1	P2	P3	P4	P5	P6	P7
Zero Shot							
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 5: 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.

clubbed pants on fire and false into one label that is false Table 1

First, we need to clean, test and val set first so that we can make use of the dataset for the experiment, then we will move on to train the dataset.

Fields in the dataset for instance: Id, label, speaker, claim, evidence, source, speaker, claim_data, etc given in the provided json file. I added "leaked" and "annotator prediction" for you to fill in. Label Description: 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.

B.1 Problem with the current dataset

The dataset contains the claim and corresponding evidence that supports or refutes the claim. We have some leakage in our dataset. Leakage means, there are evidences in the dataset which are giving away the information about the label of the corresponding claim. The evidence may contain the definition of the label, some direct intuition about the label, or the label itself.

What needs to be done:

Remove our ruling section if exists.

Remove the sentence that contains a label or label definition.

Remove sentences that directly give away the information about the label.

Remove redundant conclusions by the annotator if they repeat information from the previous section or can be inferred from prior content.

Mark which evidence needed changes in the "leaked" field by writing yes/no. Give your predicted label in the "annotator prediction" field.

Examples with strike-throughs helped annotators understand what to remove. We met weekly and re-annotate if needed.

"label": "true"
"claim": "At nearly 19 million people, the population of Florida is larger than all the earlier primary and caucus states combined."
"evidence": "Gov. Rick scott rallied republican activists at florida's presidency 5 straw poll with an argument for the state's supremacy in choosing the party's presidential contender.none will have a greater impact on the selection of the nominee than our own primary in the sunshine state, scott told a crowd of 3,500 on sept. 24, 2011.while other primaries or caucuses might be earlier, he said, florida's population and diversity set it apart.at nearly 19 million people, the population of florida is larger than all the earlier primary and caucus states combined, he said.the republican national committee allows just iowa, new hampshire, south carolina and nevada to vote in february 2012 without penalty.florida has yet to choose its primary date. But state lawmakers would like to see it as early as possible, saying it better reflects the country than the four early states and should play an agenda-setting role."
"speaker": "Rick Scott"
"claim_date": "9/24/2011"
"source": "speech"
"factchecker": "Becky Bowers"
"factcheck_date": "9/27/2011"
"factcheck_analysis_link": "https://www.politifact.com/factchecks/2011/sep/27/rick-scott/gov-rick-scotts-primary-math-florida-has-more-peop/"

Figure 2: A true instance from the dataset.

"label": "mostly-true"
"claim": "The failings in our civil service are encouraged by a system that makes it very difficult to fire someone even for gross misconduct."
"evidence": "Sen. John mccain, the arizona republican, overstates the problem of removing federal employees for poor performance, but not by much, according experts who examine federal work rules. It is perhaps not a surprise that a union offical disputes mccain's use of the incompetent federal worker cliche. Procedures do exist to remove workers from their jobs, and many people do get fired. But it takes a long time, according to the outside experts who follow such issues closely. Mccain wisely faults not an individual but a system. That puts him on pretty solid ground, where even a study by the federal government had difficulty finding supervisors who had attempted to take action against poorly performing employees."
"speaker": "John McCain"
"claim_date": "3/21/2007"
"source": "other"
"factchecker": "Angie Drobnic Holan"
"factcheck_date": "9/1/2007"
"factcheck_analysis_link": "https://www.politifact.com/factchecks/2007/sep/01/john-mccain/you-can-fire-federal-workers-but-its-tough/"

Figure 3: A mostly true instance from the dataset.

"label": "half-true"
"claim": "21-million Americans could have a four-year college scholarship for the money we've squandered in Iraq. 7.6-million teachers could have been hired last year if we weren't squandering this money."
"evidence": "Former U.S. Sen. Mike gravel attacked the iraq war during a recent debate by highlighting the increasing costs. Stop and think, he said at howard university on june 28, 2007. When he's talking about the money we're squandering, 21-million americans could have a four-year college scholarship for the money we've squandered in iraq. 7.6-million teachers could have been hired last year if we weren't squandering this money. Gravel's campaign staff didn't respond to numerous requests for documentation supporting those numbers. They couldn't even say how much they think the iraq war costs. The college board puts the average cost of tuition for a four-year public university in 2006 at $5,836. Do the math: the sum exceeds $490.2-billion, much higher than even the highest estimate. The U.S. Department of education reports the average teacher salary was $47,750 in 2005, the most recent year available. That produces a total of $363-billion, well below the lowest estimate. The congressional budget office conservatively estimates the entire bill for the iraq war since 2001 is $413-billion."
"speaker": "Mike Gravel"
"claim_date": "6/28/2007"
"source": "other"
"factchecker": "John Frank"
"factcheck_date": "9/20/2007"
"factcheck_analysis_link":
"https://www.politifact.com/factchecks/2007/sep/20/mike-gravel/hes-high-then-hes-low/"

Figure 4: A half true instance from the dataset.

"label": "mostly-false"
"claim": "Photo shows a semi-truck that crashed with a Chevy pickup that cut in front of it."
"evidence": "An unnerving photo of a vehicle crumpled under a semi-truck is being shared on social media with a warning: the next time you decide to cut in front of that 80,000 lb semi, remember: this was once a 4-door chevy pickup. In september 2016, wsb-tv, a news station in atlanta, aired images from a crash involving four tractor-trailers on interstate 20 in carroll county, georgia. Georgia state patrol said at the time that a tractor-trailer ran into the back of a second tractor-trailer, according to the station. The second tractor-trailer then drove over a silver pickup truck, crushing it, and ran into a third tractor-trailer. The third tractor trailer then hit a fourth one. The person driving the pickup and a passenger were killed in the crash. The atlanta journal-constitution reported the same narrative. The deadly chain reaction started when a tractor-trailer headed eastbound struck a second tractor-trailer, which then struck the silver pickup truck, killing the driver and passenger, the newspaper said. The photo suggests this is what happens to smaller vehicles that cut in front of big trucks on the highway. But this was no ordinary collison; it involved multiple vehicles that are not all pictured."
"speaker": "Viral image"
"claim_date": "6/15/2021"
"source": "social_media"
"factchecker": "Ciara O'Rourke"
"factcheck_date": "6/21/2021"
"factcheck_analysis_link": "https://www.politifact.com/factchecks/2021/jun/21/viral-image/crash-photo-doesnt-show-vehicle-cut-front-semi-tru/"

Figure 5: A mostly false instance from the dataset.

"label": "false"

"claim": "A 2022 video shows Ukrainian and Russian soldiers face to face."

"evidence": "Footage of soldiers firing shots into the air as hundreds of unarmed people march toward an airbase in belbek, crimea, is being shared on tiktok as russia invades ukraine. Ukrainian and russian soldiers face off in big battle border, one post sharing the footage wrote. #ukrainian and #ryssland soldiers face to face, another post said. This footage was posted over 12 times on tiktok and viewed on the platform more than 20 million times as of feb. 25. Bbc news turkey shared the footage on youtube on march 4, 2014. According to the bbc article, the video depicts pro-russian troops who seized an airbase firing warning shots to prevent some 300 unarmed ukrainian soldiers from approaching. The tense standoff occurred as russia annexed crimea in 2014."

"speaker": "TikTok posts"

"claim_date": "2/25/2022"

"source": "blog"

"factchecker": "Yacob Reyes"

"factcheck_date": "2/25/2022"

"factcheck_analysis_link":"https://www.politifact.com/factchecks/2022/feb/25/tiktok-posts/video-standoff-between-soldiers-ukraine-2014/"

Figure 6: A false instance from the dataset.