MPCG: Multi-Round Persona-Conditioned Generation for Modeling the Evolution of Misinformation with LLMs

Anonymous ACL submission

Abstract

Misinformation evolves as it spreads, shifting in language, framing, and moral emphasis to adapt to new audiences. However, current misinformation detection approaches implicitly assume that misinformation is static. We introduce **MPCG**, a multi-round, persona-conditioned framework that simulates how claims are iteratively reinterpreted by agents with distinct ideological perspectives. Our approach uses an uncensored large language model (LLM) to generate persona-specific claims across multiple rounds, conditioning each generation on outputs from the previous round, enabling the study of misinformation evolution. We evaluate the generated claims through human and LLM-based annotations, cognitive effort metrics (readability, perplexity), emotion evocation metrics (sentiment analysis, morality), clustering, and downstream classification. Results show strong agreement between human and GPT-4o-mini annotations, with higher divergence in fluency judgments. Generated claims require greater cognitive effort than the original claims and consistently reflect persona-aligned emotional and moral framing. Clustering and cosine similarity analyses confirm semantic drift across rounds while preserving topical coherence. Classification results reveal that commonly used misinformation detectors experience macro-F1 performance drops of up to 50%. The code is available at https://anonymous.4open.science/r/anonymous-repo-62D1.

1 Introduction

Misinformation remains a persistent societal threat, influencing our lives in many ways. Although the Automated Fact Checking (AFC) community has made significant strides through specialized datasets (Vlachos and Riedel, 2014; Thorne et al., 2018), improving explainability (Wang and Shu, 2023), and integrating intent features (Wang et al., 2024a), misinformation remains difficult to contain.

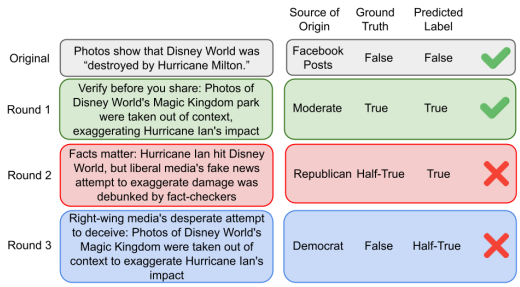

Figure 1: An illustration of how misinformation evolves across perspectives. As each persona reinterprets the original claim, static AFC systems progressively fail to classify the transformed variants, highlighting their limitations against evolving misinformation.

During the COVID-19 pandemic, Dr. Graham Walker, an emergency physician in San Francisco, left X (formerly Twitter) when the platform abandoned its COVID-19 misinformation policy[1]. His experience reflects how efforts to combat misinformation are undermined by the rapid amplification and evolution of misinformation through social media platforms (Vosoughi et al., 2018).

While modern misinformation research focuses on verifying the veracity of claims (Simeone et al., 2024), most AFC approaches implicitly assume that misinformation is static. These systems are trained and evaluated on fixed-claim datasets (Wang, 2017; Thorne et al., 2018; Schlichtkrull et al., 2023), overlooking how misinformation can persist and evolve (Bragazzi and Garbarino, 2024). As shown in Figure 1, static AFC systems progressively fail to identify these evolved variants. In reality, misinformation is dynamic: it evolves in language, framing, and moral emphasis, evades detection, and continues resonating with target audiences. This evolving nature undermines current

[1] https://str.sg/wyzo

AFC methods, making it crucial to model not only static claims but also their potential transformations.

Recent works have explored misinformation generation to pollute data in question answering systems (Pan et al., 2023), annotate claims based on selected evidence (Bussotti et al., 2024), and disguise fake news by restyling to evade fake news detectors (Wu et al., 2024). However, these are largely one-shot generation methods and fail to model how misinformation evolves ideologically. Likewise, existing AFC datasets that are derived from fact-checking websites such as PolitiFact[2] and Snopes[3] focus on verifying individual claims, without annotations capturing their variations tailored to specific audiences.

To our knowledge, limited work has explored how claims evolve through iterative reinterpretations by ideologically distinct agents. Existing LLM-based generation approaches are typically one-shot and lack mechanisms to simulate semantic and stylistic mutation across perspectives. In contrast, persona conditioning provides a structured way to simulate belief-driven reframing, while multi-round generation enables the modeling of misinformation transformation, both critical in understanding how misinformation evolves and persists.

To address this, we propose **MPCG** (**M**ulti-round **P**ersona-**C**onditioned **G**eneration), a framework that simulates misinformation evolution by iteratively reframing claims through different ideological personas. In each round, an uncensored LLM generates a new claim based on a target persona, the original and previous generated claims. This cumulative setup enables the modeling of misinformation evolution across different ideological perspectives while maintaining topic coherence.

Our main contributions are:

- We formally introduce a new task of multi-round claim generation where LLMs simulate how misinformation dynamically adapts across ideological viewpoints. It addresses a key limitation in current fact verification: the assumption that misinformation is static.
- We propose an interpretable generation framework that simulates the iterative misinformation transformation through role-playing agents. By conditioning on both prior claims and ideological personas, our method produces realistic, persona-aligned misinformation variants.
- We conduct extensive experiments encompassing human and LLM-based assessments, cognitive and emotional metrics, semantic drift analysis, and downstream detection robustness. The evaluation results demonstrate the framework's effectiveness in stress-testing current misinformation detection systems.

[2] https://www.politifact.com/
[3] https://www.snopes.com/

2 Related Works

2.1 Claim Generation with LLMs

Claim generation was first introduced as the task of producing claims from information extracted from Wikipedia using human annotators (Thorne et al., 2018). Originally motivated by data scarcity for fact verification, this approach has evolved into automated approaches. Encoder-based transformer models such as BART (Lewis et al., 2020) have been used to convert question-answer pairs into claims (Pan et al., 2021) and generate scientific claims (Wright et al., 2022). More recently, decoder-based transformer models such as LLMs have been used to generate claims based on selected evidence (Bussotti et al., 2024).

However, current claim generation frameworks do not account for how claims can be reinterpreted when expressed by individuals with different backgrounds. As misinformation spreads, each iteration subtly reshapes the structure of the claims while preserving the main topic of the original claim and its sources. Our work addresses this gap by proposing a multi-round, persona-conditioned claim generation framework, where claims are iteratively interpreted and reconstructed by role-playing agents based on their assigned personas and previously generated claims. This design enables us to model how misinformation transforms over time through social reinterpretations.

2.2 Role-Playing with LLMs

Role-playing refers to the act of aligning LLMs with specific personas or characters (Chen et al., 2025) to simulate distinct behaviors or viewpoints. Common implementations include fine-tuning open-source models with role-playing datasets (Wang et al., 2024b) or prompting models with well-crafted profiles, often referred to as personas (Wang et al., 2024c). In misinformation research, role-playing with LLMs has been used to simu-

late social media environments. Applications include generating synthetic comments through user-to-user interactions (Wan et al., 2024) and simulating the spread of rumors (Hu et al., 2025) and fake news (Liu et al., 2024).

To our knowledge, few studies have applied role-playing specifically for misinformation claim generation. Our work extends this line of research by leveraging multi-round, persona-conditioned generation to analyze how misinformation evolves across different perspectives. This setup provides a structured and interpretable way to study the forms that misinformation takes as it spreads, offering a new direction for misinformation generation research.

3 Role-Playing Claim Generation Framework

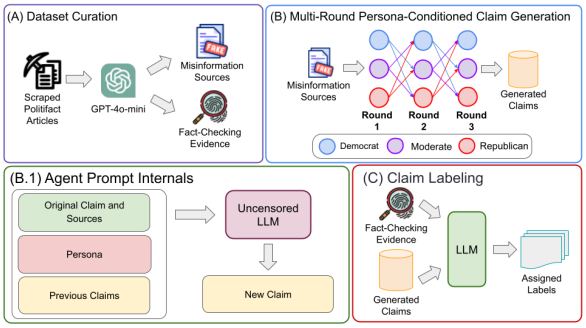

Figure 2: Overview of the **MPCG** Framework

3.1 Overview

We introduce **MPCG**, a framework designed to simulate misinformation evolution for claims. As illustrated in Figure 2, **MPCG** operates in three stages:

1. **Dataset Curation:** Scrape PolitiFact articles and use GPT-4o-mini to extract both **Misinformation Sources** and **Fact-Checking Evidence**.
2. **Multi-Round Persona-Conditioned Claim Generation:** Generate persona-aligned claims over three rounds using a structured LLM pipeline, conditioning each generation on the original claim and prior outputs to simulate misinformation evolution.
3. **Claim Labeling:** Annotate each generated claim with veracity labels (`True`, `Half-True`, `False`) using a structured LLM pipeline for downstream evaluation.

3.2 Problem Definition

Given an original claim C_0 authored by an individual CO, a set of contextual sources S, a sequence of personas $P_1, P_2, \ldots, P_k$ where each persona is a tuple (R_k, D_k) representing a role and its description, and optionally a set of previously generated claims $C_{<k}$, the goal is to generate a sequence of claims $C_1, C_2, \ldots, C_k$. Each claim C_k should reflect the viewpoint of persona P_k. We define the generation function G as:

$$C_k = G(C_{<k}, C_0, CO, S, R_k, D_k) \quad (1)$$

The function G is implemented using a multi-step prompting pipeline applied to an uncensored LLM, which Instructs the model through structured instructions to simulate persona-conditioned interpretation. The generated claims are not necessarily factually accurate, as the personas may introduce bias, exaggeration, or misleading information.

3.3 Persona Curation and Setup

We defined three personas based on the American political spectrum: `Democrat`, `Republican` and `Moderate`. These roles allow us to analyze how claims evolve across different perspectives. For each role, we curated detailed role descriptions to generate claims that align with how these roles are perceived in society. Additional details are described in Appendix A.

3.4 Dataset Curation

Motivation To ensure grounded claim generation and evaluation, our framework curate a custom dataset that addresses limitations in existing fact-checking datasets such as LIAR (Wang, 2017), LIAR-PLUS (Alhindi et al., 2018) and AVeriTeC (Schlichtkrull et al., 2023), which lack both background context and evidence.

Data Source We collect 22,408 articles from PolitiFact[4], a reputable fact-checking source in English. Each article includes a highlighted claim, background context, evidence, cited sources, and a final veracity label. Table 1 shows the raw dataset distribution up to 17 March 2025.

Data Creation Most articles follow a consistent format: an introduction presenting the claim and its background, a transition sentence marking the beginning of the debunking phase, followed by detailed debunking process and conclude with a final

[4] `https://www.politifact.com`

Label	Count	Label	Count
True	2263	Mostly False	3407
Half-True	3431	False	7010
Mostly True	3187	Pants on Fire	3110
Total			**22408**

Table 1: Raw data annotation counts from PolitiFact.

label. Due to the variability in article length and writing style, rule-based extraction proved unreliable.

Inspired by prior work (Chatrath et al., 2024), we used GPT-4o-mini to extract two components from each article using the "Our Sources" section:

- **Misinformation Sources:** Background context supporting the claim.
- **Fact-Checking Evidence:** Evidence used to verify or debunk the claim.

The complete annotation prompt is provided in Appendix D

Data Formatting and Verification The annotated outputs are cleaned and formatted. The PolitiFact veracity labels (`True, Mostly True, Half True, Mostly False, False, Pants on Fire`) are consolidated into three categories, `True`, `Half True` and `False` as shown in Table 2. This consolidation was necessary as our misinformation detectors could not distinguish subtle label differences during initial testing.

True	Half-True	False	Total
2263	6618	13527	22408

Table 2: Raw data annotations statistics after label combination

Categories	Count	Ratio (%)
No Issues	36	72.0
Contaminated Sources	9	18.0
Poor Extraction	5	10.0

Table 3: Manual verification results on 50 samples.

To evaluate the annotation quality, we manually reviewed 50 of our annotated samples. Each sample was assessed for extraction accuracy and categorized into one of three groups: `No Issues`, `Contaminated Sources`, and `Poor Extraction`. `No Issues` indicates satisfactory annotation quality, `Contaminated Sources` refers to cases where **Misinformation Sources** contain debunking statements, and `Poor Extraction` indicates low quality outputs for both **Misinformation Sources** and **Fact-Checking Evidence**. In many cases, the extracted content covered only a subset of the listed sources, likely due to the capabilities of GPT-4o-mini. Despite these limitations, the extraction quality for **Fact-Checking Evidence** was generally accurate. Overall, the annotations were deemed sufficient for our framework. Table 3 summarizes the distribution of these findings.

License and Terms of Use. The PolitiFact data used in this work was collected from publicly accessible fact-check articles through our custom web scraper. We acknowledge the terms of use provided by PolitiFact [5] and confirm that the data was collected solely for non-commercial academic research purposes.

3.5 Multi-Round Persona-Conditioned Claim Generation

MPCG simulates misinformation evolution by generating persona-specific claims over three rounds using an uncensored LLM as shown in Figure 2. In Round 1, each agent A generates a claim C_k based on the original claim C_0, its author CO, assigned persona P, and contextual sources S. In Round 2 and 3, previously generated claims $C_{<k}$ are included to model how misinformation might evolve through reinterpretation by ideologically distinct agents.

Generation is performed using **Llama-3.1-8B-Lexi-Uncensored-V2**, an uncensored **LLaMA-3.1-8B-Instruct** model provided by OrengUteng in HuggingFace [6]. This model was selected following preliminary experiments with the standard LLaMA-3.1-8B-Instruct model, which frequently rejected prompts due to its safety alignment mechanisms. Each round is implemented through a structured five-step prompting pipeline executed within a single content window:

1. **Source Reasoning Prompt**: Instructs the model to analyze and reason with the original claim C_0, its sources S, and available previous claims $C_{<k}$ from the perspective of the assigned persona P.
2. **Claim Generation Prompt**: Instructs the model to generate a new 20-word claim based

[5] https://www.politifact.com/copyright/

[6] https://huggingface.co/Orenguteng/Llama-3.1-8B-Lexi-Uncensored-V2

on its reasoning.

3. **Intent Generation Prompt**: Instructs the model to state its intent when generating the claim.
4. **Explanation Prompt**: Instructs the model to generate an explanation based on the claim.
5. **Formatting Prompt**: Request the model to provide a response in JSON format.

This prompting design ensures that each generated claim remains topically grounded with the original claim while reflecting the rhetorical and ideological biases of the assigned persona. The prompts can be found in Appendix B

3.6 Claim Labeling

To enable downstream classification, each generated claim is assigned a veracity label: `True`, `Half-True`, `False` using the provided evidence E as shown in Figure 2. We automate this process using a structured labeling pipeline using **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024). The pipeline consists of three prompts executed within a single content window:

1. **Evidence Analysis**: Guides the model in analyzing the generated claim C_k by comparing it with evidence.
2. **Label Assignment**: Asks the model to assign the appropriate label and provide a confidence score.
3. **Formatting and Label Selection**: Asks the model to select the label with the highest confidence score and return it in JSON format.

All outputs are stored in JSON for our downstream classification task. The prompts can be found in Appendix C

4 Experiments

We evaluated the effectiveness of **MPCG** by addressing the following key research questions:

1. **RQ1: Human-Level Misinformation Quality.** Can **MPCG** generate persona-conditioned misinformation that aligns with human-level quality in terms of role-playing consistency, content relevance, fluency, factuality, and veracity assignment?
2. **RQ2: Linguistic and Moral Characteristics.** What linguistic, emotional, and moral features do the generated claims exhibit, and how do these features evolve across rounds?
3. **RQ3: Impact on Classifier Robustness.** How does multi-round claim evolution affect the accuracy and robustness of existing misinformation classifiers?
4. **RQ4: Role of Contextual Grounding.** How important are background sources and persona descriptions in shaping the quality and diversity of the generated claims?

4.1 Dataset

To support claim generation and downstream classification tasks, we curate our own dataset as mentioned in Section 3.4. Table 4 presents the dataset statistics after label consolidation and class balancing. The final dataset contains 6,789 samples, evenly distributed across three classes: `True`, `Half-True`, `False`. The dataset is split into **Train, Dev, Test** with 80%/10%/10% ratio. The `Test` set is used for claim generation in our framework, while the `Train` and `Dev` sets are used to finetune the encoder-based misinformation classifiers.

Dataset Type	True	Half-True	False	Total
Train	1811	1811	1811	5433
Dev	226	226	226	678
Test	226	226	226	678

Table 4: Final dataset distribution after label consolidation and balancing

4.2 Evaluation Setup and Metrics

All experiments are performed using an NVIDIA A100 GPU on Google Colab and GPT-4o-mini. Generating 10,170 claims and labeling them took about 2 days. We evaluate the generated claims using three complementary evaluations.

Human and GPT-4o-mini Evaluation We conduct a questionnaire-based evaluation with 30 university graduates familiar with American politics. Annotators rate the generated claims based on role-playing consistency, content relevance, fluency, and factuality using a 5-point Likert scale. They are tasked with assigning veracity labels (`True`, `Half-True`, `False`) based on the provided evidence. The same task is performed with GPT-4o-mini. To quantify agreement between human and GPT-4o-mini responses across all rating dimensions, we compute the binned Jensen-Shannon Divergence (JSD) (Menéndez et al., 1997; Elangovan et al., 2025) using `jensenshannon` provided by SciPy (Virtanen et al., 2020). Additional details are provided in Appendix E.

Claim Analysis We analyze the linguistic and emotional features of these generated claims using metrics from previous work (Carrasco-Farré, 2022). These metrics are grouped into three categories: **Cognitive Effort**, **Emotion Evocation**, and **Clustering**. **Cognitive Effort** measures the processing difficulty of the claim using readability and perplexity (Carrasco-Farré, 2022). We measure readability using Flesch-Kincaid Grade Level (FKGL) score (Kincaid et al., 1975) via TextStat[7] and perplexity using GPT-2 via HuggingFace[8]. Perplexity indirectly measures lexical diversity through text quality (Tevet and Berant, 2021). **Emotion Evocation** measures the emotional appeal of the claim using sentiment analysis and morality (Carrasco-Farré, 2022). Sentiment is measured via the `sentiment-analysis` pipeline provided by HuggingFace[9] using `cardiffnlp/twitter-roberta-base-sentiment` (Barbieri et al., 2020). Morality is analyzed using MoralBERT (Preniqi et al., 2024) which measures the morality of a given text based on ten moral foundations. Additional morality details are provided in Appendix F. **Clustering** measures the semantic deviations between the generated claims and their original claims using `all-MiniLM-L6-v2` model provided by SBERT (Reimers and Gurevych, 2019), HDBSCAN (Malzer and Baum, 2020) with `min_cluster_size` of 5 and UMAP (McInnes et al., 2018) with a random state of 42 in its default settings.

Classification We evaluate the impact of our generated claims on downstream tasks using classification with commonly used encoder-based and decoder-based models in misinformation detection. We measure the macro precision, recall, and F-1 scores using `precision_score`, `recall_score`, `f1_score` provided by Scikit-Learn (Pedregosa et al., 2011). For encoder-based models, we fine-tune BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and DeBERTa-v3 (He et al., 2023) and their large variants using the HuggingFace Trainer[10] framework and our training dataset stated in Section 4.1. For decoder-based models, we use LLaMA 3.1-8B Instruct and GPT-4o-mini to label these claims via zero-shot, few-shot, zero-shot chain-of-thought (CoT) (Wei et al., 2022) and few-shot CoT prompting strategies. The prompts, fine-tuning approaches and additional details are provided in Appendix G.

[7] https://pypi.org/project/textstat/

[8] https://huggingface.co/docs/transformers/en/perplexity

[9] https://huggingface.co/blog/sentiment-analysis-python

[10] https://huggingface.co/docs/transformers/en/main_classes/trainer

5 Results and Discussion

We evaluate **MPCG** across the following stages: **Original** refers to the original PolitiFact claims; **Round 1, 2, 3** correspond to the generated claims using our framework in Figure 2. The evaluation uses 678 **Original**, 2034 **Round 1**, and 4068 **Round 2 and 3** claims.

5.1 Human and GPT-4o-mini Evaluation

Question	JSD
Role-Playing Consistency (Q1)	0.178
Content Relevance (Q2)	0.174
Fluency (Q3)	0.296
Factuality (Q4)	0.195
Label Assignment (Q5)	0.113

Table 5: Jensen Shannon Divergence (JSD) scores for 363 human and GPT-4o-mini evaluations. Lower is better.

Table 5 shows strong alignment between human and GPT-4o-mini ratings for most dimensions, with higher divergence in fluency. GPT-4o-mini favors "Excellent" while humans tend to select "Good", reflecting a potential judgment bias (Chen et al., 2024). Despite this, both rate fluency highly overall, indicating general fluency of generated claims.

5.2 Claim Analysis

Round	Persona	Median	Q1	Q3	IQR
1	Democrat	14.04	11.96	16.05	4.09
1	Moderate	14.07	12.31	16.04	3.73
1	Republican	14.37	12.48	16.20	3.72
2	Democrat	14.63	12.57	16.76	4.19
2	Moderate	14.60	12.44	16.46	4.02
2	Republican	14.81	12.86	16.76	3.90
3	Democrat	14.93	12.86	16.88	4.02
3	Moderate	14.93	12.84	16.99	4.15
3	Republican	14.64	12.83	16.88	4.05
–	Original	9.14	6.92	11.95	5.03

Table 6: Flesch-Kincaid Grade Level (FKGL) scores for original and generated claims across all rounds. Higher scores indicate more syntactically complex text.

Cognitive Effort Table 6 shows that generated claims are syntactically more complex (median FKGL = 14.0 to 14.9) than the original (median

Round	Role	Median	Q1	Q3	IQR
1	Democrat	45.33	29.87	72.56	42.70
1	Moderate	51.48	34.83	76.18	41.35
1	Republican	47.52	31.00	71.07	40.07
2	Democrat	42.40	29.69	66.29	36.60
2	Moderate	50.16	35.50	74.92	39.42
2	Republican	48.12	33.38	75.32	41.94
3	Democrat	44.67	30.90	71.04	40.14
3	Moderate	49.82	34.52	79.26	44.74
3	Republican	48.76	33.84	74.67	40.83
–	Original	56.97	32.78	107.23	74.45

Table 7: Perplexity scores for original and generated claims across all rounds. Lower scores indicate less lexical diversity and higher word-level predictability.

Role (Round)	MFT Dimension	Avg Score	SE
Republican (Round 3)	Authority	0.1061	0.0066
Democrat (Round 2)	Betrayal	0.0482	0.0044
Democrat (Round 3)	Care	0.1832	0.0063
Democrat (Round 1)	Cheating	0.1834	0.0123
Original	Degradation	0.0721	0.0043
Democrat (Round 3)	Fairness	0.2643	0.0106
Original	Harm	0.1085	0.0090
Moderate (Round 3)	Loyalty	0.0179	0.0026
Original	Purity	0.0066	0.0025
Republican (Round 3)	Subversion	0.0135	0.0024

Table 9: Morality scores across original and generated claims based on Moral Foundations Theory (MFT). Higher scores indicate stronger moral framing expressed in the analyzed claims.

FKGL = 9.1). Table 7 shows that they are also less lexically diverse (median Perplexity = 43.4 to 51.5) when compared to the original (median Perplexity = 56.9). These two results indicate that the generated claims require a higher education level to comprehend but use a consistent vocabulary set.

Claim Source	Round	Negative	Neutral	Positive
Original	-	281	**350**	47
Democrat	1	**365**	179	134
Moderate	1	152	**399**	127
Republican	1	**319**	195	164
Democrat	2	**598**	435	323
Moderate	2	300	**759**	297
Republican	2	**540**	403	413
Democrat	3	**547**	436	373
Moderate	3	222	**760**	374
Republican	3	**491**	457	408

Table 8: Distribution of sentiments for original and generated claims across rounds and personas. Bolded values indicate the majority sentiment class within each group.

Emotion Evocation Table 8 shows Democrat and Republicans produce more negative claims, while Moderates remain mostly neutral. Table 9 shows Democrats emphasize care, fairness, cheating, and betrayal, while Republicans emphasize authority and subversion, aligning with previous work where liberals rely heavily on harm and fairness while conservatives rely more on authority (Day et al., 2014). In contrast, Moderates generate more neutral claims while emphasizing loyalty, likely reflecting a downplaying or a neutralizing strategy during the generation process. These results indicate that our framework can generate claims that are morally and sentimentally aligned with the personas, poised to resonate with its intended audience.

Clustering Figure 3 shows the clusters formed by a sample of 300 claims using SBERT, UMAP and HDBSCAN. Each color corresponds to a unique PolitiFact URL, and each shape represents a different generation round. Our clustering results show that most generated claims remain semantically close to their respective original claims, suggesting strong topic coherence throughout the generation process. However, a subset of generated claims that deviates significantly from their original claims due to the shifts in framing. These results indicate that misinformation can evolve stylistically, changing its tone, emphasis and perspective while preserving topic alignment.

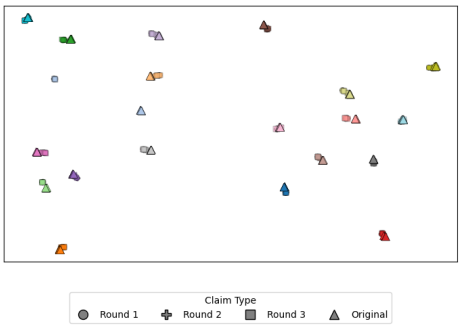

Figure 3: Semantic clusters of Round 1 generated claims (Circle), Round 2 generated claims (Plus), Round 3 generated claims (Square), and the original claims (Triangle). Each color represents a group of claims associated with the same original PolitiFact URL.

5.3 Classification Robustness

Table 10 shows that DeBERTa V3$_{Large}$ achieves the highest classification performance in the original claims (macro F1 = 0.72), but all models suffer significant drops on generated claims: 45% to 50% for encoder-based models and between 17% to 46%

Model	Original			Round 1			Round 2			Round 3		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BERT_{Base}	0.64	0.65	0.64	0.33	0.33	0.33	0.37	0.38	0.37	0.36	0.36	0.36
BERT_{Large}	0.67	0.65	0.66	0.37	0.36	0.36	0.38	0.38	0.38	0.37	0.36	0.35
RoBERTa_{Base}	0.68	0.67	0.67	0.36	0.35	0.35	0.39	0.38	0.38	0.37	0.37	0.37
RoBERTa_{Large}	0.71	0.71	0.71	0.37	0.36	0.36	0.40	0.39	0.39	0.39	0.38	0.38
DeBERTa V3_{Base}	0.70	0.69	0.70	0.39	0.39	0.38	0.37	0.37	0.36	0.38	0.37	0.37
$\text{DeBERTa V3}_{Large}$	**0.73**	**0.71**	**0.72**	0.39	0.37	0.36	0.40	0.39	0.38	0.41	0.39	0.38
LLaMA 3.1 8B Instruct (Zero Shot)	0.62	0.62	0.61	**0.45**	**0.45**	**0.44**	**0.46**	**0.46**	**0.46**	**0.44**	**0.44**	**0.43**
LLaMA 3.1 8B Instruct (Zero Shot CoT)	0.61	0.55	0.52	0.50	0.45	0.43	0.46	0.44	0.41	0.43	0.40	0.37
LLaMA 3.1 8B Instruct (Few Shot)	0.61	0.57	0.56	0.43	0.42	0.40	0.45	0.45	0.42	0.43	0.41	0.38
LLaMA 3.1 8B Instruct (Few Shot CoT)	0.62	0.53	0.51	0.44	0.41	0.37	0.43	0.42	0.38	0.44	0.40	0.35
GPT-4o-mini (Zero Shot)	0.69	0.68	0.68	0.47	0.42	0.42	0.51	0.44	0.43	0.49	0.43	0.41
GPT-4o-mini (Zero Shot CoT)	0.71	0.65	0.65	0.47	0.40	0.37	0.53	0.44	0.42	0.51	0.41	0.38
GPT-4o-mini (Few Shot)	0.72	0.68	0.69	0.45	0.39	0.37	0.50	0.40	0.37	0.51	0.41	0.38
GPT-4o-mini (Few Shot CoT)	0.73	0.67	0.68	0.48	0.40	0.37	0.49	0.40	0.37	0.50	0.40	0.36

Table 10: Macro Average Precision (P), Recall (R) and F1 scores for each model across grouped claims: Original, Round 1, Round 2, and Round 3. Bolded values indicate highest macro F1 within each group of claims.

Datasets	Count	Avg Cosine Similarity
Original → Round 1	2034	0.6445
Original → Round 2	4068	0.6248
Original → Round 3	4068	0.6177
Round 1 → Round 2	4068	0.6992
Round 2 → Round 3	4068	0.6978

Table 11: Average Cosine Similarities for each dataset combinations

for decoder-based models. This aligns with prior findings that stylistic perturbations can reduce fake news detectors F1 score performance up to 38% (Wu et al., 2024).

To explain the performance plateau in later rounds, we analyze the average cosine similarities in Table 11. The average cosine similarities with the original claims steadily decline from Round 1 to Round 3, indicating an incremental semantic drift. However, adjacent rounds (Round 1 → and Round 2, and Round 2 → and Round 3) maintain high similarity (approx 0.70), suggesting that each generation introduces only small shifts. These results indicate that while evolving misinformation can reduce the accuracies of these models, they still show some robustness when claims remain semantically close.

6 Conclusion and Future Work

In this work, we introduce a new task: multi-round claim generation, where LLMs simulate how misinformation dynamically adapts across ideological viewpoints. We propose **MPCG**, a novel framework that models misinformation evolution through iterative generation and role-playing agents while preserving topic coherence.

Human and GPT-4o-mini evaluations show strong alignment in role-playing consistency, relevance, fluency, and factuality, indicating that the generated claims can mimic human level misinformation. Claim analysis reveals that the generated claims require higher cognitive effort and exhibit persona-aligned sentiment and moral framing, suggesting their potential to influence targeted audiences. Clustering and cosine similarity analyses further confirm that claims evolve stylistically and semantically over rounds, while retaining topic alignment.

Classification results indicate that standard misinformation detectors suffer performance degradation on Round 1 claims, with performance plateau in later rounds. This highlights the models' robustness to semantically similar claims and vulnerability to stylistic shifts.

These findings emphasize the evolving nature of misinformation and the importance of modeling its progression. Our framework provides a foundation for stress-testing AFC systems, similar to load testing in software engineering, to help develop more resilient detection methods.

Future work includes developing standardized automatic metrics to evaluate generation quality, reducing reliance on subjective human and LLM assessments prone to judgment bias (Chen et al., 2024) and human uncertainty (Elangovan et al., 2025). Additionally, extending this framework to multilingual and multimodal settings can broaden its applicability and provide insights into how misinformation evolves in different settings.

Limitations

We discover several limitations throughout our work. First, the dataset curation process as discussed in Section 4.1, can be affected by the quality and variability of both PolitiFact articles and the annotation process using GPT-4o-mini. Although GPT-4o-mini enables scalable and consistent annotations, it may introduce inaccuracies in the `Misinformation Sources` and `Fact-Checking Evidence`. The annotation quality is sensitive to prompt design and models, which may affect reproducibility across setups.

Second, our framework relies on curated personas to simulate different ideological perspectives. Although these personas are grounded in established typologies, they might not fully represent the current state of the roles. Additionally, the typologies used in our work were published in 2021 which does not accurately reflect the current trends of our selected personas. Another probable issue is that it might be difficult to curate non-generic personas such as creating a persona that does not have established typologies.

Third, the generation pipeline of our framework depends on a single uncensored LLMs, `LLaMA-3.1-8B-Lexi-Uncensored-V2`. Although this model was selected due to the stated reasons mentioned in Section 3.5, the results can differ when we use different uncensored models with the same configurations.

Next, our evaluation metrics used in our thesis may limit the interpretability and novelty of our findings. While human evaluation was conducted with care and precision, the number of annotators and our questionnaire design may not fully reflect the broader or more diverse perspectives. Similarly, the ground truth labels used for our classifications are based on LLM-generated labels rather than gold-standard annotations from experts which could introduce compounding errors in metric-based assessments.

Finally, there are some limitations with our claim analysis metrics. First, the Flesch-Kincaid Grade Level is not a robust metric as it is based on fairly simplistic text statistics which is exploitable to get good scores (Tanprasert and Kauchak, 2021). Second, perplexity indirectly measures the lexical diversity of a text, making it a poor assessment. Third, the sentiment analysis and morality metrics interpretations are due to subjectivity as they rely on pre-trained classifiers whose predictions may not reflect real-world dynamics.

Ethical Considerations

This work involves generating synthetic misinformation content using uncensored LLMs, which presents some ethical risks. While the objective is to study how misinformation evolves, we acknowledge that generating such content may be misused and cause intended harm. To mitigate this, we do not release any generated claims. Instead, we provide the generation code and framework configuration, allowing researchers to replicate the methodology under controlled settings.

Our persona definitions are based on publicly available sources and are intended to reflect real-world discourse, not to reinforce stereotypes. However, these personas are U.S centric and derived from typologies established in 2021 which may not reflect the current state of communities that share the same typologies. As such, the findings of our study should not be generalized to non-U.S perspectives.

The dataset used contains only publicly available information disclosed in PolitiFact articles. No personally identifiable information is included, and we do not infer any protected attributes such as race, gender, and ethnicity. All data is used for non-commercial academic research.

The outputs of our framework are synthetic and not intended for public deployment. Nonetheless, there is a potential for unintended misuse, including the interpretation of generated claims as real texts. Researchers may find such tools useful for modeling the evolution of misinformation but should exercise cautious in avoiding reinforcing false narratives.

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.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Nicola Luigi Bragazzi and Sergio Garbarino. 2024. Understanding and combating misinformation: An evolutionary perspective. *JMIR Infodemiology*, 4:e65521.

Jean-Flavien Bussotti, Luca Ragazzi, Giacomo Frisoni, Gianluca Moro, and Paolo Papotti. 2024. Unknown claims: Generation of fact-checking training examples from unstructured and structured data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12105–12122, Miami, Florida, USA. Association for Computational Linguistics.

Carlos Carrasco-Farré. 2022. The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions - Humanities and Social Sciences Communications — nature.com. https://www.nature.com/articles/s41599-022-01174-9.

Veronica Chatrath, Marcelo Lotif, and Shaina Raza. 2024. Fact or fiction? can llms be reliable annotators for political truths? *Preprint*, arXiv:2411.05775.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2025. The oscars of ai theater: A survey on role-playing with language models. *Preprint*, arXiv:2407.11484.

Martin V Day, Susan T Fiske, Emily L Downing, and Thomas E Trail. 2014. Shifting liberal and conservative attitudes using moral foundations theory. *Personality and Social Psychology Bulletin*, 40(12):1559–1573.

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.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, et al. 2025. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and llm-as-a-judge. *Preprint*, arXiv:2410.03775.

Anthony Fowler, Seth J Hill, Jeffrey B Lewis, Chris Tausanovitch, Vavreck, et al. 2023. Moderates. *American Political Science Review*, 117(2):643–660.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Kadian, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Tianrui Hu, Dimitrios Liakopoulos, Xiwen Wei, Radu Marculescu, and Neeraja J. Yadwadkar. 2025. Simulating rumor spreading in social networks using llm agents. *Preprint*, arXiv:2502.01450.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, et al. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. *Preprint*, arXiv:2403.09498.

Claudia Malzer and Marcus Baum. 2020. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, page 223–228. IEEE.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. 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 476–483, Online. Association for Computational Linguistics.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages

1389–1403, Singapore. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Vjosa Preniqi, Iacopo Ghinassi, Julia Ive, Charalampos Saitis, and Kyriaki Kalimeri. 2024. Moralbert: A fine-tuned language model for capturing moral values in social discussions. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*, GoodIT '24, page 433–442, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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*.

Michael Simeone, Kristy Roschke, and Shawn Walker. 2024. Evolutionary biology as a frontier for research on misinformation. *Politics and the Life Sciences*, page 1–3.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

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.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

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.

Soroush Vosoughi, Deb K. Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359:1146 – 1151.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. DELL: Generating reactions and explanations for LLM-based misinformation detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2637–2667, Bangkok, Thailand. Association for Computational Linguistics.

Bing Wang, Ximing Li, Changchun Li, Bo Fu, Songwen Pei, et al. 2024a. Why misinformation is created? detecting them by integrating intent features. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2304–2314, New York, NY, USA. Association for Computing Machinery.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6288–6304, Singapore. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024b. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. 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.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024c. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 257–279, Mexico City, Mexico. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference*

on Neural Information Processing Systems, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 3367–3378, New York, NY, USA. Association for Computing Machinery.

A Persona Curation

To ensure that each persona reflects a socially grounded and ideologically representative viewpoint, we establish role descriptions based on multiple sources.

For `Democrat` and `Republican` role descriptions, their definitions were derived from the Wikipedia version of Beyond Red vs. Blue: The Political Typology 2021 from Pew Research Center [11], which describes the American political spectrum in 2021 modeled by Pew Research Center. We specifically referenced the Democratic Coalition and the Republican Coalition typologies for their definitions. Due to the variability of the length of our claim's sources, these references were summarized to respect token limitations. Additionally, we incorporate the definitions from Merriam-Webster [12] which provides a high-level definition of the role.

The definition for `Moderate` is derived from Wikipedia's Political Moderate page [13] which provides a high-level definition for this role. Furthermore, we also incorporate the characteristics of a `Moderate` from American Political Science Review (Fowler et al., 2023) which offers a scholarly perspective for this role.

The final role descriptions used for our framework are listed below:

A.1 Democrat

- Younger liberal voters that are skeptical of the political system and both major political parties. They believe that the American political system unfairly favors powerful interests, and about half say that the government is wasteful and inefficient. They are more likely to say that no political candidate represents their political views and least likely to say that there is a "great deal of difference" between the parties.

- Older voters that are economically liberal and socially moderate who support higher taxes and expansion of the social safety net as well as stronger military policy. They also see violent crime as a "very big" national problem, to oppose increased immigration, and to say that people being too easily offended is a major problem.

- Highly liberal voters who are loyal to the Democratic Party and are more likely than other groups to seek compromise and to hold an optimistic view of society.

- Younger highly liberal voters who believe that the scope of government should "greatly expand" and that the institutions of the United States need to be "completely rebuilt" to combat racism. They are the most likely group to say that there are countries better than the United States, that the American military should be reduced, that fossil fuels should be phased out, and that the existence of billionaires is bad for society.

- A member of one of the two major political parties in the U.S. that is usually associated with government regulation of business, finance, and industry, with federally funded educational and social services, with separation of church and state, with support for abortion rights, affirmative action, gun control, and policies and laws that protect and support the rights of workers and minorities, and with internationalism and multilateralism in foreign policy.

A.2 Republican

- Highly conservative and highly religious voters who generally support school prayer and military over diplomacy while generally oppose legalized abortion and same-sex marriage. They are more likely to claim that the United States "stands above all other countries in the world" and that illegal immigration is a "very big national problem", known to be staunch pro-Israel supporters, are more likely to reject the concept of white privilege and to agree that white Americans face more discrimination than African Americans and people of color.

[11] https://en.wikipedia.org/wiki/Pew_Research_Center_political_typology
[12] https://www.merriam-webster.com/
[13] https://en.wikipedia.org/wiki/Political_moderate

- Conservative voters that emphasize pro-business views, international trade and small government who hold moderate views on immigration and race than other groups within the Republican coalition.
- Highly conservative anti-immigrant voters that oppose the role of government and big businesses in American society. They are more likely to believe that the number of legal immigrants should decrease and that the decreasing proportion of white Americans is bad for society. They are also more likely to support raising taxes on the rich.
- Younger voters that lean conservative on economic and race issues but lean moderate on social issues. They are more likely to support diplomacy over military strength, legalized marijuana, legalized abortion and "openness to people from all over the world".
- A member of one of the two major political parties in the United States that is usually associated with reduced taxation, with limited government regulation of business, finance, industry, education, and policing, with strong national defense, and with opposition to abortion, affirmative action, gun control, and policies and laws that are viewed as challenging traditional social and family hierarchies and structure.

A.3 Moderate

- An ideological category which designates a rejection of radical or extreme views, especially in regard to politics and religion.
- Someone occupying any mainstream position to avoid extreme views.
- Often described as politically unsophisticated, uninformed, or ideologically innocent, secretly partisan, ideologically cross-pressured, or extreme, with patterns of attitudes poorly described by a single ideological dimension

B Claim Generation Prompt Templates

B.1 Source Reasoning without Previous Claims Prompt

You are a role. A role is described as follows:
{role_definition}
Your task is to analyze and reason about the original claim the original claim owner, and its context from your role's perspective.

Original Claim Owner: {claim_owner}
Original Claim: {original_claim}
Original Claim Context (Sources): {sources}

Your reasoning should address:
1) Role Interpretation:
- How does your role perceive this claim?
- Does your role agree, disagree, or reframe it differently?
- Would your role see this claim as valid, misleading, or needing modification?

2) Key Emphasis & Framing:
- What parts of this claim would your role emphasize or downplay?
- How does your role's beliefs or agenda influence what details are most important?

Summarize your reasoning clearly and concisely before proceeding to modify the claim.
Ensure that your reasoning remains consistent with the role's values and agenda.

B.2 Source Reasoning with Previous Claims Prompt

You are a {role}. A {role} is described as follows: {role_definition}

Your task is to analyze and reason about the original claim, the original claim owner, and its context from your role's perspective.
You should also consider how previous claims have evolved from the original claim and whether your role agrees, reframes, or challenges them.

Original Claim Owner: {claim_owner}
Original Claim: {original_claim}
Original Claim Context (Sources): {sources}
Previous Claims (from previous rounds): {previous_claims}

Your reasoning should address:
1) Role Interpretation:
- How does your role perceive these claims?
- Does your role agree, disagree, or reframe it differently?
- Would your role see these claim as valid, misleading, or needing modification?

2) Key Emphasis & Framing:
- What parts of these claim would your role emphasize or downplay?
- How does your role's beliefs or agenda influence what details are most important?

Summarize your reasoning clearly and concisely before proceeding to modify these claims.
Ensure that your reasoning remains consistent with the role's values and agenda.

B.3 Claim Generation Prompt

Based on your role's reasoning, generate a persuasive yet factually coherent claim.

Ensure that:
- The claim remains aligned with the original claim and sources.
- The wording, tone, emphasis, and persuasion reflect your role's beliefs.
- The claim is natural, engaging, clear, concise, and no more than 20 words.

Modified Claim:

B.4 Intent Generation Prompt

State your intent when generating this claim based on your role.

Consider:
- What message is your role trying to convey with this claim?
- What reaction does your role want to provoke in the audience?
- Does this claim aim to inform, persuade, create doubt, or reinforce a belief?
- How does your role's ideology shape the claim's purpose?

Ensure the response is written in a single, coherent sentence.

B.5 Explanation Prompt

Provide a structured explanation of the modified claim.

Your response should include:
- How was the claim modified from the original?
- Why does the modification align with your role's beliefs and perspective?
- How does the claim remain factually coherent while reflecting your role's emphasis?
- What effect is the claim intended to have on the audience?

Ensure the explanation flows naturally as a single, concise sentence.

B.6 Formatting Prompt

Return the Claim, Intent, and Explanation in JSON Format.

Ensure that:
- The Claim remains aligned with the original claim and sources.
- The Intent clearly defines the purpose of the claim.
- The Explanation justifies the claim's modification while maintaining logical consistency.

Format the response as follows:

```json
{{
 "Claim": "<Modified claim>",
 "Intent": "<Purpose of the claim>",
 "Explanation": "<How and why the claim was modified>"
 }}
```

C Claim Labeling Prompt Templates

C.1 Evidence Analysis Prompt

You are a fact-checking assistant. Your task is to analyze the claim and compare it with the provided evidence.

Claim:
"claim"

Evidence:
"evidence"

Instructions:
1. Carefully analyze whether the claim is fully, partially, or not supported by the evidence.
2. Identify specific factual elements in the claim that are supported or contradicted.
3. Note any missing context, exaggerations, or misleading aspects of the claim.
4. Do not make assumptions beyond what the evidence explicitly states.

C.2 Label Assignment Prompt

Based on your factual analysis, assign the appropriate label.

Label Definitions:
- True: A statement is fully accurate.
- Half-True: A statement that conveys only part of the truth, especially one used deliberately in order to mislead someone.
- False: A statement is inaccurate or contradicted by evidence.

Instructions:
- Avoid assumptions beyond what the analysis states.
- Ensure consistency between the label and reasoning.
- Provide your confidence score for all of the labels.

C.3 Formatting and Label Selection Prompt

Select the label based on the highest confidence score and provide an explanation on your factual analysis.

Output Format (JSON)
```json
{{
"Label": "<True / Half-True / False>",
"Explanation": "<Short justification referencing your factual analysis>"
}}
```

D Dataset Annotation Prompt

You are a fact-checking annotator trained to extract and categorize information from the given "Article" based on the "Original Sources", "Original Claim" and "Original Claim Label".

Your task is to extract "Misinformation Sources" and "Fact-Checking Evidence" from the given "Article" based on the "Original Sources", "Original Claim" and "Original Claim Label".

A fact-checking annotator is a role that helps to assign a truth value to a claim made in a particular context.

Consider the following in your evaluation:
Definitions:

- Politifact 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.
  - Barely True: 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 / Pants-on-Fire: The statement is not accurate and makes a ridiculous claim.
- Misinformation Sources:
  - Information that is incorrect or misleading but is typically spread without malicious intent.
  - This can include errors, misinterpretations, or contextually misleading statements that can be amplified or misunderstood when shared.
  - Along with each misleading statement, you must specify the source of origin, who made the statement, where it originated, and its description.
- Fact-Checking Evidence:
  - Verified information from reliable sources that is used to assess the veracity of a claim suspected of being misinformation.
  - This can include statements from experts, official documentation, government officials, or trustworthy organizations that clarify misunderstandings or provide factual context to refute misleading claims.
  - Along with each verified information, you must include any supporting information or transitioning information that support this information.
- Rating Sentence:
  - A sentence that indicate the final evaluation of the claim's accuracy based on "Politifact Labels"
- Transition Sentence:
  - A sentence that introduces the "Fact-Checking Evidence" section's topic or argument, contain logical connectors or indicate a shift in focus, tone, or evidence from the "Misinformation Sources" section.

  - It is not the same as "Rating Sentence"

- "Social Media Flag" Sentences:
  - Sentences that contains these sentences that indicates partnership with social media companies
    - "Social Media Flag" sentences examples:
      - "Read more about PolitiFact's partnership with Meta"
      - "Read more about PolitiFact's partnership with TikTok"

- "Question" Sentences:
  - Sentences that is structured as a question and clearly indicates a transition between "Misinformation Sources" and "Fact-Checking Evidence".
    - "Question" sentences examples
      - Is Hochul right? Have 732,000 jobs have been created since she became governor in late summer 2021?

- "But" Sentences:
  - Sentences that starts with "But" and clearly indicates a transition between "Misinformation Sources" and "Fact-Checking Evidence"
    - "But" sentences examples:
      - But there's no record Trump, the president-elect, ever said those words. These viral videos use old footage with what appears to be fake audio generated by artificial intelligence.
      - But these social media posts are wrong. Haley was born in South Carolina and meets the U.S. Constitution's requirements to run for president.

- "Action" Sentences:
  - Sentences that indicates an action and a transition between "Misinformation Sources" and "Fact-Checking Evidence". It is not the same as "Rating Sentence".
    - "Action" sentences examples:
      - For this fact-check, we examined only Biden's comment about wages and inflation.
      - We decided to look into how the university system has been funded in recent years.

- "Reasoning" Sentences:
  - Sentences that does not have clear transition indicators, but are indicated as a "Transition Sentence" when the whole article is considered.
    - "Reasoning" sentences examples:
      - PolitiFact New Jersey found Doherty is mostly right: when a student is enrolled in the federally supported lunch program, they are designated as at risk.

Please remember these definitions.

There are two tasks that you will need to do.

Preprocessing Task:

- Read the whole article
- Identify the "Transition Sentence" from "Misinformation Sources" to "Fact-Checking Evidence" that are similar to "Social Media Flag" Sentences, "But" Sentences, "Question" Sentences, "Action" Sentences and "Reasoning" Sentences.

- Use the "Transition Sentence" to split the article into "Misinformation Sources Chunk" and "Fact-Checking Evidence Chunk" without any modifications.
- Retrieve the sentences from "Misinformation Sources Chunk" that are originated from "Original Sources" without any modifications as "Misinformation Sources".
  - Additional sentences that describe the main sentences must be included as well.
- Retrieve the sentences from "Fact-Checking Evidence Chunk" that are originated from "Original Sources" without any modifications as "Fact-Checking Evidence".
  - Additional sentences that describe the main sentences must be included as well.

Cleanup Task:

- Do not include the "Transition Sentence" and "Rating Sentence" as an output for "Misinformation Sources" and "Fact-Checking Evidence".
- The sentences in "Misinformation Sources" and "Fact-Checking Evidence" must be sorted based on the sentence ordering in the article.
- Combine the sentences in "Misinformation Sources" based on the "Article" and "Original Sources".
- Combine the sentences in "Fact-Checking Evidence" based on the "Article" and "Original Sources".
- Return the "Misinformation Sources", "Fact-Checking Evidence", "Transition Sentence" and "Explanation" in JSON.

Please perform the task as stated.

Important rules:

- You must consider all sentences in the article.
- You are not allowed to rephrase or modify any texts from the article.
- There cannot be two identical texts in both "Misinformation Sources" and "Fact-Checking Evidence".
- "Misinformation Sources" sentences must include the person or source that mentioned the sentence.
- Sentences that contain "Politifact Labels" regardless of its form must be excluded from the output.
- "Transition Sentence" must not be included in "Misinformation Sources" and "Fact-Checking Evidence".
- "Rating Sentence" must not be the same as "Transition Sentence".

Please follow the rules strictly.

"Original Claim":
`{original_claim}`

"Original Claim Label":
{original_claim_label}

"Original Sources":
{our_sources}

"Article":
{article}

E Human and GPT-4o-mini Evaluation

We conducted a structured evaluation using both GPT-4o-mini and human annotators to assess the quality of our generated claims, as described in Section 4.2. We recruited 30 university graduates familiar with American politics that are based in Singapore, Malaysia and United States via personal academic networks. No financial compensation was provided as participation was voluntary and required minimal time commitment, and all participants consented without objection.

Annotators were instructed to complete a set of questionnaires to the best of their abilities. The evaluation was conducted anonymously via Google Forms. Annotators were told that their responses would be used for academic research and that no personal information would be collected. Each form contained a generated claim, its associated persona, paraphrased content. The contents were paraphrased with GPT-4o-mini due to the length of the sources, evidence, and role descriptions.

Annotators rated each claim on four dimensions using a 5-point Likert scale where 1 is the lowest and 5 is the highest.

1. **Role-Playing Consistency:** How well does the Claim align with the Role's beliefs and intention?
2. **Content Relevance:** How relevant is the Claim compared to the provided sources?
3. **Fluency:** How fluent the Claim is in terms of grammar, clarity, and readability?
4. **Factuality:** How factually correct is the Claim?

In addition to the Likert ratings, the annotators were asked to assign a veracity label to each claim based on the provided evidence using our consolidated labels scheme: `True`, `Half-True`, or `False`. We evaluated a sample of 363 unique claims from all three rounds of role-playing generation. These claims were randomly selected from the generated set to ensure fairness and diversity. Below we present an example of our questionnaire.

E.1 Question 1: Role-Playing Consistency

How well does the Claim align with the Role's beliefs and intention?

Role: Democrat

Claim: Rhode Islanders deserve an honest vote on same-sex marriage, not flawed polls that misrepresent their true opinions.

Role Description:
- Encompasses younger liberal voters who are disillusioned with the political system, viewing it as biased towards powerful interests, while also including older economically liberal voters who support higher taxes and a stronger social safety net.
- Includes highly liberal members who are loyal to the Democratic Party and advocate for significant governmental reform to address social issues, emphasizing the need to combat racism

and reduce military presence.
- Represents a political orientation focused on government regulation, social justice, individual rights, and international cooperation, often advocating for policies like abortion rights, affirmative action, and worker protections.

Intent: To persuade the audience to question the validity of biased polls and to advocate for a more inclusive and democratic process that respects the rights of same-sex couples.

* 5 - Perfectly Consistent: The claim fully aligns with the role's beliefs, tone, and intent.

* 4 - Mostly Consistent: The claim follows the role and intent but may miss small details.

* 3 - Somewhat Consistent: The claim partly aligns but lacks key points or misrepresents intent.

* 2 - Mostly Inconsistent: The claim contradicts some role beliefs but has a weak connection.

* 1 - Completely Inconsistent: The claim opposes or has no connection to the role.

E.2 Question 2: Content Relevance

How relevant is the Claim compared to the provided sources and previous claims?

Role: Democrat

Claim: Rhode Islanders deserve an honest vote on same-sex marriage, not flawed polls that misrepresent their true opinions.

Sources:
- Over 80% of Rhode Islanders support the opportunity to vote on the issue, according to multiple surveys conducted by Quest Research.
- Three surveys, each with around 400 participants and a margin of error of 20 ± 4%, were analyzed, including findings from a June 2009 poll.
- Concerns were raised about question design in these polls, particularly due to phrasing like "backroom politicians," which may skew responses.

Previous Claims:
- Republican: Polls show a small minority of Rhode Islanders, not 80%, support same-sex marriage, and most want government out of it

* 5 - Perfectly Relevant: The claim fully integrates key facts from sources and previous claims.

* 4 - Mostly Relevant: The claim follows sources or previous claims but misses small details.

* 3 - Somewhat Relevant: The claim mentions sources or previous claims but misinterprets or lacks connections.

* 2 - Weakly Relevant: The claim has a weak or indirect connection to the sources or previous claims.

* 1 - Completely Irrelevant: The claim does not relate to any sources or previous claims.

E.3 Question 3: Fluency

How factually correct is the Claim?

Claim: Rhode Islanders deserve an honest vote on same-sex marriage, not flawed polls that misrepresent their true opinions.

Sources:
- Over 80% of Rhode Islanders support the opportunity to vote on the issue, according to multiple surveys conducted by Quest Research.
- Three surveys, each with around 400 participants and a margin of error of 20 ± 4%, were analyzed, including findings from a June 2009 poll.
- Concerns were raised about question design in these polls, particularly due to phrasing like "backroom politicians," which may skew responses.

Previous Claims:
- Republican: Polls show a small minority of Rhode Islanders, not 80%, support same-sex marriage, and most want government out of it

Evidence:
- Question wording by interest groups can significantly influence poll results, potentially skewing public perception (John Geer).
- Phrasing polls in a way that reflects biases can lead to inflated support for certain views (John Geer).
- Neutral questions tend to reveal stronger support for rights, like gay marriage, with independent pollsters estimating that around 80% of responses favor equality when phrased objectively.

* 5 - Excellent: Clear, well-written, and grammatically perfect.

* 4 - Good: Mostly correct, with minor errors that do not affect readability.

* 3 - Adequate: Readable but has noticeable errors or awkward phrasing.

* 2 - Poor: Contains multiple errors that make it harder to understand.

* 1 - Very Poor: Frequent errors make the claim difficult to comprehend.

E.4 Question 4: Factuality

How factually correct is the Claim?

Claim: Rhode Islanders deserve an honest vote on same-sex marriage, not flawed polls that misrepresent their true opinions.

* 5 - Completely Accurate: Fully factual, with no misleading parts or missing context.

* 4 - Mostly Accurate: Mostly factual, with small mistakes or missing details that don't change the meaning.

* 3 - Partially Accurate: Some parts are true, but others are misleading or missing key facts.

* 2 - Mostly Inaccurate: Many errors or missing key details, making it misleading.

* 1 - Completely Inaccurate: Completely false or highly misleading.

E.5 Question 5: Label Assignment

Which label do you think is suitable for the Claim based on the evidence?

Claim: Rhode Islanders deserve an honest vote on same-sex marriage, not flawed polls that misrepresent their true opinions.

Evidence:
- Question wording by interest groups can significantly influence poll results, potentially skewing public perception (John Geer).
- Phrasing polls in a way that reflects biases can lead to inflated support for certain views (John Geer).
- Neutral questions tend to reveal stronger support for rights, like gay marriage, with independent pollsters estimating that around 80% of responses favor equality when phrased objectively.

* True: Fully accurate.

* Half-True: Partially accurate, lacks important details or is misleading

* False: Inaccurate

F Morality Details

Morality captures the measurement of emotions through social identity. For our work, we analyze the moral framing of generated claims using MoralBERT (Preniqi et al., 2024), a BERT (Devlin et al., 2019) model that is fine-tuned to capture moral sentiment in social discourse based on Moral Foundations Theory (MFT). MoralBERT predicts the presence of ten moral dimensions, grouped into five psychological foundations which are listed below.

- `Care/Harm` are defined as the involvement of concern for others' suffering and includes virtues like empathy and compassion.
- `Fairness/Cheating` focuses on issues of unfair treatment, inequality, and justice.
- `Loyalty/Betrayal` pertains to group obligations such as loyalty and the vigilance against betrayal.
- `Authority/Subversion` centers on social order and hierarchical responsibilities, highlighting obedience and respect.
- `Purity/Degradation` refers to relates to physical and spiritual sanctity, incorporating virtues like chastity and self-control.

G Classification Configurations

G.1 Encoder Config

As stated in Section 4.2, we finetuned BERT$_{Base}$ (110M Parameters), RoBERTa$_{Base}$ (125M Parameters), DeBERTA V3$_{base}$ (184M Parameters), BERT$_{Large}$ (340M Parameters), RoBERTa$_{Large}$ (355M Parameters), DeBERTA V3$_{Large}$ (435M Parameters) for our experiments with HuggingFace Trainer and our training dataset on Google Colab A100 (40 GB).

RoBERTa and DeBERTa fine-tuning configurations were adopted from their original papers (Liu et al., 2019; He et al., 2023) with epochs set at 5. For DeBERTA V3$_{Large}$, we used a reduced batch size of 8 due to GPU memory constraints. BERT was fine-tuned with a configuration of 5 epochs, a learning rate of 3e-5, batch size of 16, and a weight decay of 0.1. Fine-tuning each base model took approximately 2.5 hours while large models took approximately 6 hours. These estimates include training and checkpointing overhead. Below shows the TrainingArguments used to train these models.

G.1.1 BERT$_{Base}$ and BERT$_{Large}$

```
training_args = TrainingArguments(
    output_dir=SAVE_PATH,
    evaluation_strategy="epoch",
    save_strategy="epoch",
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    num_train_epochs=5,
    learning_rate=3e-5,
    weight_decay=0.1,
    logging_dir=f"{SAVE_PATH}/logs",
    load_best_model_at_end=True,
    metric_for_best_model="accuracy",
    logging_steps=1000,
    report_to="none"
)
```

G.1.2 DeBERTA V3$_{Base}$ and DeBERTA V3$_{Large}$

```
training_args = TrainingArguments(
    output_dir=SAVE_PATH,
    evaluation_strategy="epoch",
    save_strategy="epoch",
    learning_rate=3e-5,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    num_train_epochs=5,
    weight_decay=0.01,
    logging_dir=f"{SAVE_PATH}/logs",
    logging_steps=1000,
    metric_for_best_model="accuracy",
    max_grad_norm=1.0,
    warmup_steps=500,
    report_to="none"
)
```

G.1.3 RoBERTa$_{Base}$ and RoBERTa$_{Large}$

```
training_args = TrainingArguments(
    output_dir=SAVE_PATH,
    evaluation_strategy="epoch",
    save_strategy="epoch",
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    num_train_epochs=5,
    learning_rate=3e-5,
    weight_decay=0.1,
```

```
        logging_dir=f"{SAVE_PATH}/logs",
        load_best_model_at_end=True,
        metric_for_best_model="accuracy",
        logging_steps=1000,
        warmup_ratio=0.06,
        report_to="none"
    )
```

G.2 Prompts

As described in Section 4.2, we used GPT-4o-mini and LLaMA 3.1 8B Instruct to conduct our classification experiments. LLaMA 3.1 8B Instruct was run on a Google Colab A100 GPU (40 GB) with a batch size of 8, max tokens setting of 8192 and a temperature setting of 0.7, while GPT-4o-mini was accessed using its Batch API. The prompt templates used for both models are detailed below.

G.2.1 Zero-Shot

Based on the "Fact-Checking Evidence", select a "Label" from ["True", "Half-True", "False"] that is suitable for the "Claim" and provide an "Explanation".

Claim: <|claim|>

Fact-Checking Evidence:
<|fce|>

Output Format:
{{ "Label": "<True / Half-True / False>",
"Explanation": "<Short justification referencing your label choice>"
}}

G.2.2 Zero-Shot CoT

Consider the following in your evaluation:
Definitions:
- Claim:
- A statement that state or assert that something is the case, typically without providing evidence or proof
- Fact-Checking Evidence:
- Verified information from reliable sources that is used to assess the veracity of a claim suspected of being misinformation.
- This can include statements from experts, official documentation, or trustworthy organizations that clarify misunderstandings or provide factual context to refute misleading claims.

Grading Scheme:
- The scheme has three ratings, in decreasing level of truthfulness - true : The statement is accurate and there's nothing significant missing.
- half-true : The statement is partially accurate but leaves out important details or takes things out of context.
- false : The statement is not accurate.

Claim: <|claim|>

Fact-Checking Evidence:
<|fce|>

Your task is to assign an appropriate 'Label' to the 'Claim' based on its level of truthfulness, using the 'Grading Scheme'.
To determine the claim's accuracy, you must fact-check it against the 'Fact-Checking Evidence'.
After assessing the claim, you must provide a detailed 'Explanation' justifying your choice of label.
The final output must include both the 'Label' and 'Explanation' in JSON format.

Output Format:
{{ "Label": "<True / Half-True / False>",
"Explanation": "<Short justification referencing your label choice>"
}}

G.2.3 Few-Shot

Based on the "Fact-Checking Evidence", select a "Label" from ["True", "Half-True", "False"] that is suitable for the "Claim" and provide an "Explanation".

Examples:
Claim: "A proposed constitutional amendment "would allow anyone to run for a 3rd term. Including — Barack Obama."
Fact-Checking Evidence:
- fce_source_0: "But the resolution doesn't propose changing the 22nd Amendment so that former President Barack Obama — or any other president who served two consecutive terms — could run.",
- fce_source_1: "Ogles wants the amended amendment to say: 'No person shall be elected to the office of the president more than three times, nor be elected to any additional term after being elected to two consecutive terms, and no person who has held the office of the president, or acted as president, for more than two years of a term to which some other person was elected president shall be elected to the office of the president more than twice.'",
- fce_source_2: "That means former President Grover Cleveland, who died in 1908 and served two nonconsecutive presidential terms, would have been the only other former U.S. president eligible to run for reelection after serving two terms under the proposed amendment.",
- fce_source_3: "For the Constitution to be amended, Ogles' bill would need to be approved by a two-thirds vote in both the House and Senate, and then ratified by three-fourths of the states."

This is the expected output format:
{{ "Label": "False"
"Explanation": "The claim that the proposed constitutional amendment "would allow anyone to run for a 3rd term, including Barack Obama" is false. The resolution introduced by U.S. Rep. Andy Ogles does not propose changes to the 22nd Amendment to allow presidents who have served two consecutive terms, like Barack Obama, to run for a third term. Instead, it specifically states that "no person shall be elected to the office of the president more than three times" but maintains restrictions for those who have already served two consecutive terms. The proposed amendment would only allow individuals who served nonconsecutive terms, such as former President Grover Cleveland, to run again. Furthermore, amending the Constitution requires an extensive process, including approval by two-thirds of both the House and Senate, as well as ratification by three-fourths of the states, making such a change highly unlikely. Therefore, the evidence provided directly contradicts the claim and clarifies the intent and scope of the proposed resolution." }}

Claim: "Wisconsin makes it more difficult for its citizens to vote than almost any state in the nation. "
Fact-Checking Evidence:
- fce_source_0: "When asked to back up the claim, Common Cause Wisconsin Executive Director Jay Heck said he was pulling from the expertise of UW-Madison political science professor Barry Burden, who wrote previously for The Observatory that Wisconsin's voter ID law is one of the strictest in the country. (Source: Barry Burden via The Observatory, April 16, 2024)",
- fce_source_1: "Burden repeated it in an email to PolitiFact Wisconsin, writing that 'Wisconsin demands more than nearly all of the other states' when it comes to getting a ballot. (Source: Email exchange with Barry Burden, UW-Madison)",
- fce_source_2: "The National Conference of State Legislatures lists Wisconsin as one of just nine states with 'strict' photo ID laws used to identify voters. (Source: National Conference of State Legislatures, accessed Jan. 18, 2025)",
- fce_source_3: "Besides the strict voter ID law, Wisconsin has stringent rules for voter registration drives, does not have automatic voter registration, does not have preregistration for young voters, does not allow all voters to join a permanent absentee ballot list, does not consider Election Day a public holiday and does require a witness' signature on absentee ballots. (Source: National Conference of State Legislatures and Movement Advocacy Project, accessed Jan. 18, 2025)",
- fce_source_4: "Wisconsin ranks as the fifth-hardest state to vote in the country, according to the Cost of Voting Index. (Source: Cost of Voting Index, accessed Jan. 15, 2025)",
- fce_source_5: "In 1996, Wisconsin was ranked the fourth-most-accessible state in the nation for voters because it was among very few that offered same-day voter registration. (Source: Cost of Voting Index and Michael Pomante, accessed Jan. 15, 2025)",
- fce_source_6: "The state's most dramatic drop occurred between 2011 and 2015, when former Republican Gov. Scott Walker signed the voter ID requirement into law and it took effect despite a swarm of lawsuits seeking to knock it down. (Source: CBS News, 'Walker signs photo ID requirement into law,' May 25, 2011)"

This is the expected output format:
{{
"Label": "True"
"Explanation": "The claim that Wisconsin makes it more difficult for its citizens to vote than almost any other state is supported by comprehensive evidence from credible sources. These include statements by political science expert Barry Burden, data from the National Conference of State Legislatures highlighting Wisconsin's strict voter ID laws, and findings from the Cost of Voting Index ranking Wisconsin as the fifth-hardest state for voting. Additional evidence points to the state's lack of measures like automatic voter registration, preregistration for young voters, and early voting, as well as its stringent requirements for absentee voting. Historical data also shows a significant decline in voter accessibility since 2011, when the voter ID law was enacted. Taken together, this evidence confirms that Wisconsin's voting laws and policies significantly hinder accessibility compared to most other states, making the claim accurate."
}}
Claim: <|claim|>
Fact-Checking Evidence:
<|fce|>

Please return this output only. This is the expected output format:
{{
"Label": "<True / Half-True / False>",
"Explanation": "<Short justification referencing your label choice>"

}}

G.2.4 Few-Shot CoT

Consider the following in your evaluation:
Definitions:
- Claim:
- A statement that state or assert that something is the case, typically without providing evidence or proof
- Fact-Checking Evidence:
- Verified information from reliable sources that is used to assess the veracity of a claim suspected of being misinformation.
- This can include statements from experts, official documentation, or trustworthy organizations that clarify misunderstandings or provide factual context to refute misleading claims.

Grading Scheme:
- The scheme has three ratings, in decreasing level of truthfulness
- true : The statement is accurate and there's nothing significant missing.
- half-true : The statement is partially accurate but leaves out important details or takes things out of context.
- false : The statement is not accurate.

Your task is to assign an appropriate 'Label' to the 'Claim' based on its level of truthfulness, using the 'Grading Scheme'.
To determine the claim's accuracy, you must fact-check it against the 'Fact-Checking Evidence'.
After assessing the claim, you must provide a detailed 'Explanation' justifying your choice of label.
The final output must include both the 'Label' and 'Explanation' in JSON format.

Examples:
Claim: "A proposed constitutional amendment "would allow anyone to run for a 3rd term. Including — Barack Obama."
Fact-Checking Evidence:
- fce_source_0: "But the resolution doesn't propose changing the 22nd Amendment so that former President Barack Obama — or any other president who served two consecutive terms — could run.",
- fce_source_1: "Ogles wants the amended amendment to say: 'No person shall be elected to the office of the president more than three times, nor be elected to any additional term after being elected to two consecutive terms, and no person who has held the office of the president, or acted as president, for more than two years of a term to which some other person was elected president shall be elected to the office of the president more than twice.'",
- fce_source_2: "That means former President Grover Cleveland, who died in 1908 and served two nonconsecutive presidential terms, would have been the only other former U.S. president eligible to run for reelection after serving two terms under the proposed amendment.",
- fce_source_3: "For the Constitution to be amended, Ogles' bill would need to be approved by a two-thirds vote in both the House and Senate, and then ratified by three-fourths of the states."

This is the expected output format:
{{
"Label": "False"
"Explanation": "The claim that the proposed constitutional amendment "would allow anyone to run for a 3rd term, including Barack Obama" is false. The resolution introduced by U.S. Rep.

Andy Ogles does not propose changes to the 22nd Amendment to allow presidents who have served two consecutive terms, like Barack Obama, to run for a third term. Instead, it specifically states that "no person shall be elected to the office of the president more than three times" but maintains restrictions for those who have already served two consecutive terms. The proposed amendment would only allow individuals who served nonconsecutive terms, such as former President Grover Cleveland, to run again. Furthermore, amending the Constitution requires an extensive process, including approval by two-thirds of both the House and Senate, as well as ratification by three-fourths of the states, making such a change highly unlikely.Therefore, the evidence provided directly contradicts the claim and clarifies the intent and scope of the proposed resolution."
}}

Claim: "Wisconsin makes it more difficult for its citizens to vote than almost any state in the nation. "
Fact-Checking Evidence:
- fce_source_0: "When asked to back up the claim, Common Cause Wisconsin Executive Director Jay Heck said he was pulling from the expertise of UW-Madison political science professor Barry Burden, who wrote previously for The Observatory that Wisconsin's voter ID law is one of the strictest in the country. (Source: Barry Burden via The Observatory, April 16, 2024)",
- fce_source_1:"Burden repeated it in an email to PolitiFact Wisconsin, writing that 'Wisconsin demands more than nearly all of the other states' when it comes to getting a ballot. (Source: Email exchange with Barry Burden, UW-Madison)",
- fce_source_2:"The National Conference of State Legislatures lists Wisconsin as one of just nine states with 'strict' photo ID laws used to identify voters. (Source: National Conference of State Legislatures, accessed Jan. 18, 2025)",
- fce_source_3:"Besides the strict voter ID law, Wisconsin has stringent rules for voter registration drives, does not have automatic voter registration, does not have preregistration for young voters, does not allow all voters to join a permanent absentee ballot list, does not consider Election Day a public holiday and does require a witness' signature on absentee ballots. (Source: National Conference of State Legislatures and Movement Advocacy Project, accessed Jan. 18, 2025)",
- fce_source_4:"Wisconsin ranks as the fifth-hardest state to vote in the country, according to the Cost of Voting Index. (Source: Cost of Voting Index, accessed Jan. 15, 2025)",
- fce_source_5:"In 1996, Wisconsin was ranked the fourth-most-accessible state in the nation for voters because it was among very few that offered same-day voter registration. (Source: Cost of Voting Index and Michael Pomante, accessed Jan. 15, 2025)",
- fce_source_6:"The state's most dramatic drop occurred between 2011 and 2015, when former Republican Gov. Scott Walker signed the voter ID requirement into law and it took effect despite a swarm of lawsuits seeking to knock it down. (Source: CBS News, 'Walker signs photo ID requirement into law,' May 25, 2011)"

This is the expected output format:
{{
"Label": "True"
"Explanation": "The claim that Wisconsin makes it more difficult for its citizens to vote than almost any other state is supported by comprehensive evidence from credible sources. These include statements by political science expert Barry Burden, data from the National Conference of State Legislatures highlighting Wisconsin's strict voter ID laws, and findings from the Cost of Voting Index ranking Wisconsin as the fifth-hardest state for voting. Additional evidence points to the state's lack of measures like automatic voter registration, preregistration for young voters, and early voting, as well as its stringent requirements for absentee voting. Historical data also shows a

significant decline in voter accessibility since 2011, when the voter ID law was enacted. Taken together, this evidence confirms that Wisconsin's voting laws and policies significantly hinder accessibility compared to most other states, making the claim accurate."
}}

Claim: <|claim|>
Fact-Checking Evidence:
<|fce|>

Please return this output only. This is the expected output format:
{{
"Label": "<True / Half-True / False>",
"Explanation": "<Short justification referencing your label choice>"
}}

H Experiment Results

H.1 Human and GPT-4o-mini Evaluation

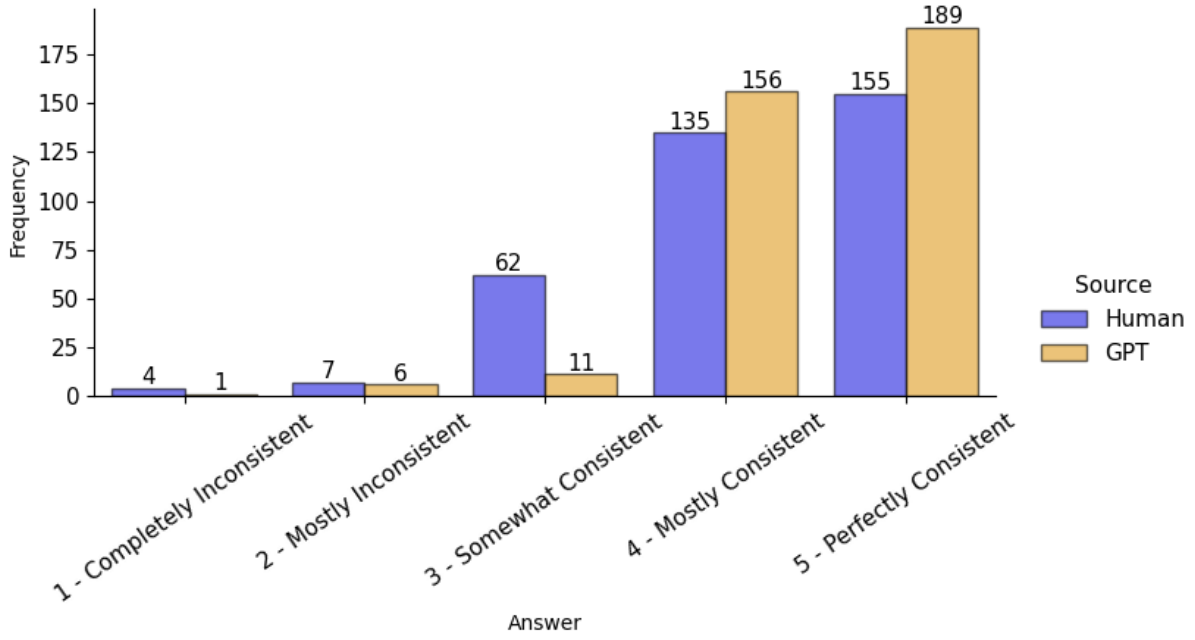

Figure 4: Role-Playing Consistency scores between human annotators and GPT-4o-mini

Figure 4 compares the Role-Playing Consistency ratings given by human annotators and GPT-4o-mini. Both raters showed similar preference, particularly for "Mostly Consistent" and "Perfectly Consistent", suggesting that our claims showed consistent role-playing effects. However, human annotators showed a slight disagreement with GPT-4o-mini where human annotators rated our generated claims "Somewhat Consistent" 62 times, compared to only 10 instances by GPT-4o-mini. This small discrepancy indicates that human raters were more likely to detect subtle inconsistencies in role consistency.

Figure 5 describes the comparison of Content Relevance scores between human annotators and GPT-4o-mini. Both annotators agree that the generated claims are relevant to their sources as indicated by the high counts of "Mostly Relevant". However, there are some disagreement in the "Somewhat Relevant" category where humans rated 91 times when compared to GPT-4o-mini at 44 times. This indicates that the human annotators are more meticulous in detecting subtleties in content relevancy. Similarly, human

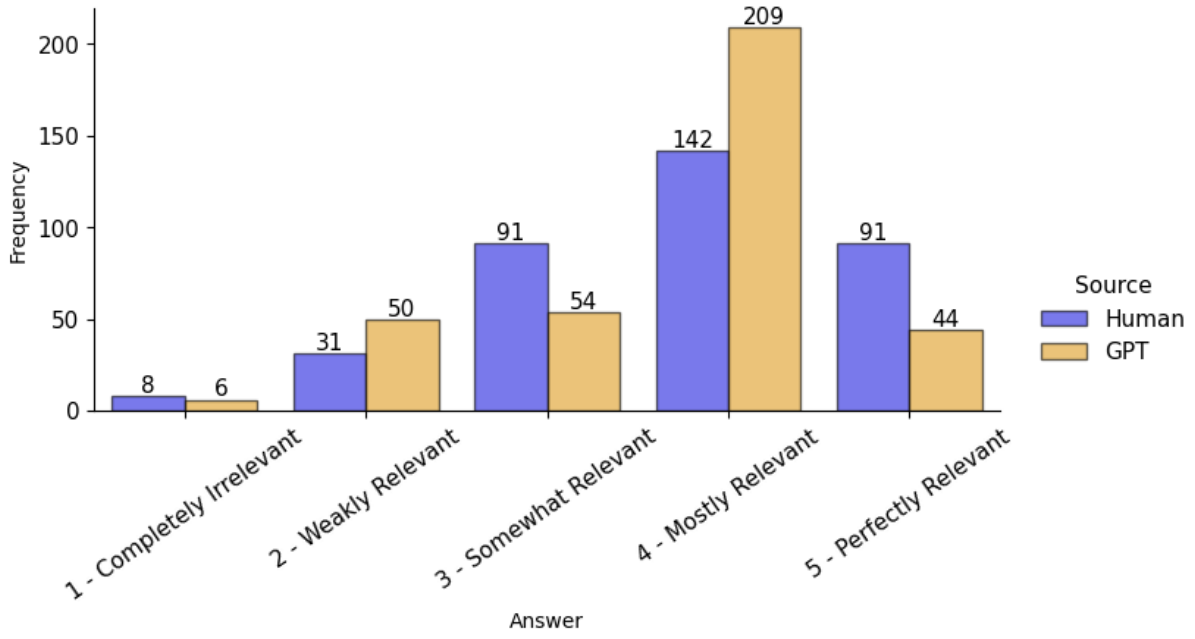

Figure 5: Content Relevance scores between human annotators and GPT-4o-mini

raters rated "Perfectly Consistent" 91 times as opposed to GPT-4o-mini at 54 times. This suggest that humans annotators based on their reasoning and prior knowledge, while GPT-4o-mini may be constrained by its training data.

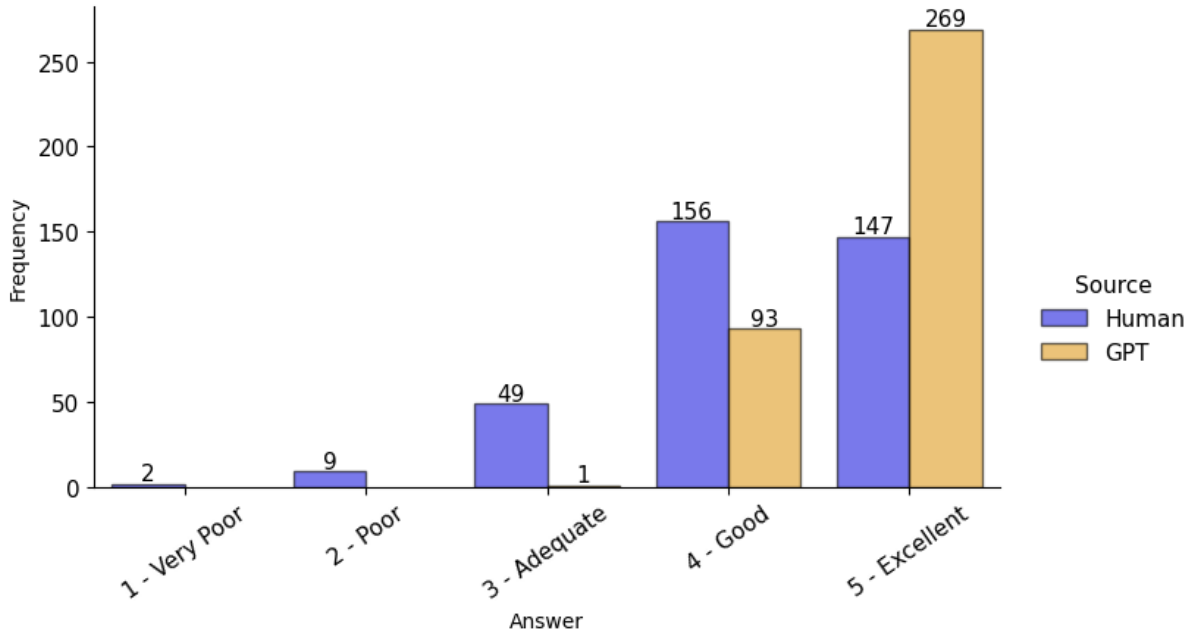

Figure 6: Fluency scores between human annotators and GPT-4o-mini

Figure 6 shows the distribution of the fluency scores between human annotators and GPT-4o-mini. Both annotators rated the majority of claims as either "Good" or "Excellent", reflecting the quality of our generated claims. However, human annotators have shown to rate 49 of our generated claims as "Adequate" where GPT-4o-mini only answered 2 for this category. This indicates that some of these structures for these claims do not show human-like fluency, while they do show that to GPT-4o-mini. Meanwhile, GPT-4o-mini rated 231 claims as "Excellent" when compared to human annotators at 147.

We suspect that the GPT-4o-mini is biased toward good answers, unlike human annotators who showed some restrictions when assessing these claims.

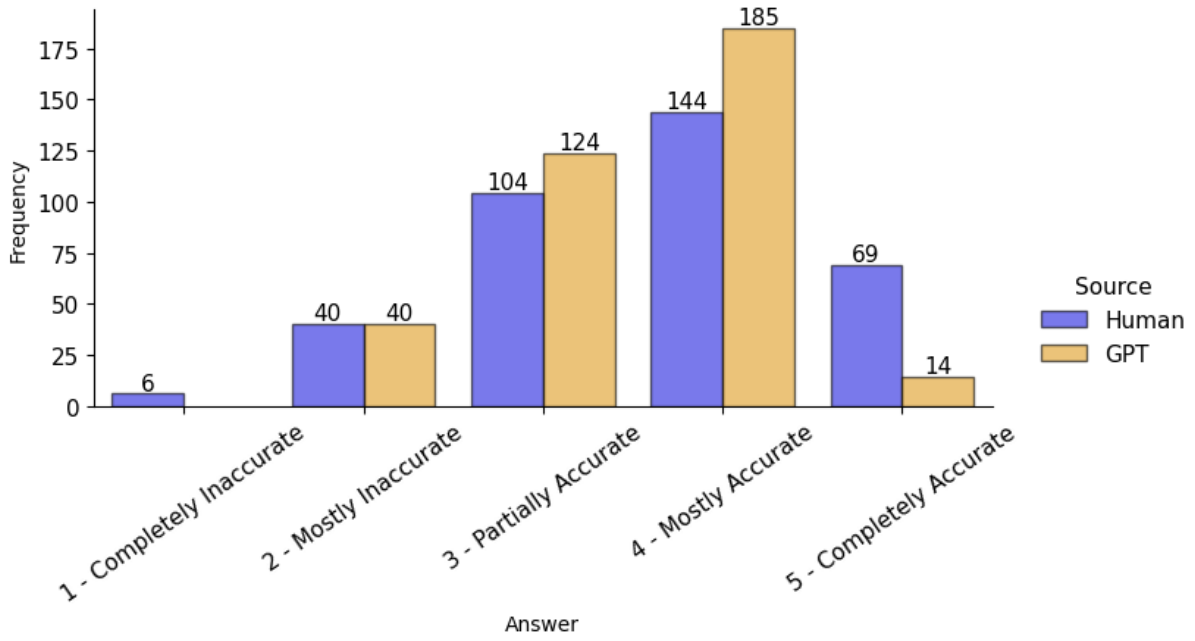

Figure 7: Factuality scores between human annotators and GPT-4o-mini

Figure 7 presents the distribution of factuality scores between GPT-4o-mini and human annotators. Both sources generally agree that our claims are factually accurate as indicated by the high "Mostly Accurate" and "Completely Accurate" counts. A notable trend here is that both annotators have rated 69 claims to be "Completely Accurate" while GPT-4o-mini rated 25 claims for the same category. This indicates that humans may have relied on their personal judgment to assess the factuality of these generated claims as opposed to GPT-4o-mini which uses its parametric knowledge. Despite this divergence, the overall similarity in distribution across the scale reflects a broad agreement between human and GPT-4o-mini.

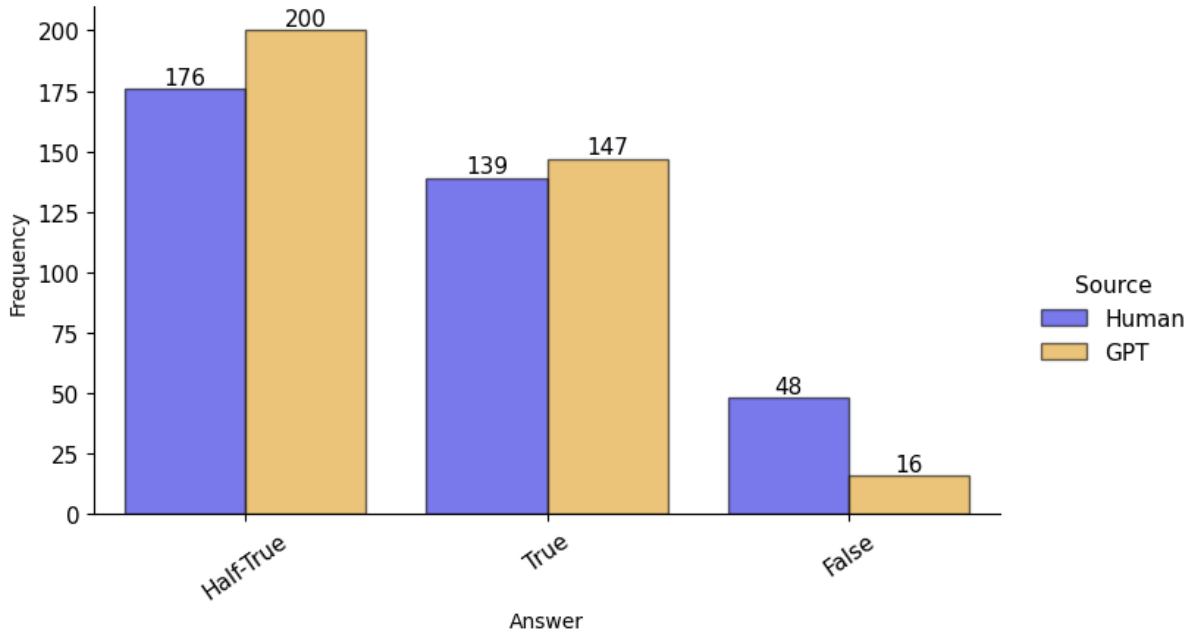

Figure 8: Label assignment between human annotators and GPT-4o-mini

Figure 8 shows the label assignment between humans and GPT-4o-mini. Both annotators labeled a large distribution of claims as "Half-True" and "True", indicating similar agreement in their assessments. A notable trend is the high 48 "False" count from human annotators when compared to GPT-4o-mini at 25. This disparity may indicate that human annotators applied more precise or stringent criteria when identifying factual inaccuracies, unlike GPT-4o-mini which may have been more lenient.