RESEARCH-ARTICLE

Personalizing LLM Responses to Combat Political Misinformation

ADIBA PROMA, University of Rochester, Rochester, NY, United States

NEELEY PATE, University of Rochester, Rochester, NY, United States

JAMES N DRUCKMAN, University of Rochester, Rochester, NY, United States

GOURAB GHOSHAL, University of Rochester, Rochester, NY, United States

EHSAN HOQUE, University of Rochester, Rochester, NY, United States

Published: 16 June 2025

Citation in BibTeX format

UMAP '25: 33rd ACM Conference on User Modeling, Adaptation and Personalization
June 16 - 19, 2025
New York City, USA

Conference Sponsors:
SIGCHI

https://doi.org/10.1145/3699682.3728349
ISBN: 9798400713132

Personalizing LLM Responses to Combat Political Misinformation

Adiba Proma
Computer Science
University of Rochester
Rochester, NY, USA
aproma@cs.rochester.edu

Neeley Pate
Computer Science
University of Rochester
Rochester, NY, USA
npate@ur.rochester.edu

James Druckman
Political Science
University of Rochester
Rochester, USA
jdruckma@UR.Rochester.edu

Gourab Ghoshal
Physics, Computer Science
University of Rochester
Rochester, NY, USA
gourab.ghoshal@gmail.com

Ehsan Hoque
Computer Science
University of Rochester
Rochester, NY, USA
mehoque@cs.rochester.edu

Abstract

Despite various efforts to tackle online misinformation, people inevitably encounter and engage with it, especially on social media platforms. Recent advances in LLMs present an opportunity to develop personalized interventions to address misinformed beliefs, and potentially offer more effective approaches than existing non-tailored methods. In this paper, we design and evaluate personalized LLM agent that can consider users' demographics and personalities to tailor responses to mitigate misinformed beliefs. Our pipeline is grounded in facts through an external Retrieval Augmented Generation (RAG) knowledge base and is able to generate diverse output as a result of the personalization, with an average cosine similarity of 0.538. Our pipeline scores an average rating of 3.99 out of 5 when evaluated by a GPT-4o-mini LLM judge for response persuasiveness. Our methods can be adapted to design similar personalized agents in other domains.

CCS Concepts

• **Human-centered computing** → **User models**; *Collaborative and social computing theory, concepts and paradigms*; • **Social and professional topics** → *User characteristics.*

Keywords

Personalization, Misinformation, LLM Agents, Persuasion, RAG systems

ACM Reference Format:
Adiba Proma, Neeley Pate, James Druckman, Gourab Ghoshal, and Ehsan Hoque. 2025. Personalizing LLM Responses to Combat Political Misinformation. In *33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '25), June 16–19, 2025, New York City, NY, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3699682.3728349

ACM ISBN 979-8-4007-1313-2/25/06
https://doi.org/10.1145/3699682.3728349

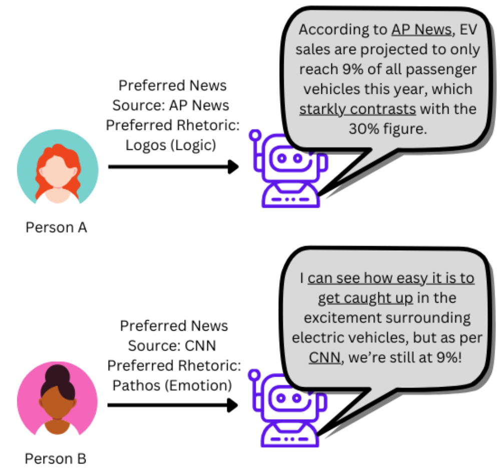


Figure 1: A high-level diagram depicting example outcomes of our proposed system. Consider the false claim – "Electric vehicles and hybrid vehicles have surged in popularity in the United States, making up 30% of all cars on the road." Assuming Person A and Person B have different personas and therefore different news source and rhetoric preferences, the proposed LLM agent would tailor responses accordingly.

1 Introduction

False news has been shown to spread six times faster than the truth on social media [38]. This rapid spread of misinformation, i.e. communication that is "false, misleading, or [based on] unsubstantiated information" [24] – can have serious real-world consequences, like influencing elections, or worsening public health crises such as COVID-19, where misinformation fueled vaccine hesitancy [6]. As generative AI tools like large language models (LLMs) and large vision-langugage models (LVLMs) become more ubiquitous, there is growing concern about misinformation getting rampant and harder to detect [5], especially on social media.

At the same time, LLMs have enhanced potential to counter misinformation [18]. LLMs have been used to detect misinformation through various fact-checking methods [26] such as integrating stance and emotion [39], knowledge distillation [41] and generating explanations to support or refute claims [40]. Yet, detecting misinformation is not sufficient on its own. Misinformation online is widespread enough that individuals will inevitably encounter it. Moreover, individuals are more likely to believe the misinformation if it aligns with their existing beliefs or resonates with them [30]. Simply informing them about the information being false has been shown to be ineffective [24]. A promising alternative approach is to generate more targeted persuasive interventions aimed at reducing belief in the encountered misinformation. This might be possible with recent advancements in generative AI, which can be used to design personalized interventions to counter misinformed beliefs – a research gap still relatively underexplored in literature.

In this paper, we explore if it is possible to counter misinformed beliefs in a personalized manner (Figure 1). For the context of our research, a claim is considered misinformation if it is demonstrably false, i.e. a claim against which there exist objective empirical contradicting evidence [12], and a misinformed belief would be believing that the claim is true. However, if the claim is in fact true, a misinformed belief would involve incorrectly believing that the claim is false. We design and evaluate personalized AI agents that can consider the user's persona and their established trust in credible news sources to support belief updates toward the truth in an empathetic manner (for example, see Figure 1). Our specific research questions are:

- RQ1: How to design an LLM agent that can provide personalized information to support or disprove a claim?
- RQ2: How well does the proposed system create diverse content, given the goal of personalization?
- RQ3: How well does the proposed system combat hallucinations, given the goal of grounding responses in truth?

To design the LLM pipeline, we first model user persona considering their demographic information and personality [29], using collaborative filtering. Given a specific statement and user persona, the pipeline is designed to pull relevant information from multiple credible sources using a retrieval-augmented generation (RAG) method. The output is ranked according to the user preference, then summarized, after which different persuasion techniques are used to convince the user regarding the veracity of the statement.

To show the efficacy of our design, we evaluate both the individual components of the pipeline as well as the final output, testing the pipeline as a whole. We investigate the effectiveness of the outputs, the diversity of all outputs when compared to each other, and the groundedness of the outputs within real news articles. Our pipeline is rated to be convincing when evaluated by an LLM judge, with an average score of 3.99 out of 5. Moreover, our pipeline generates diverse outputs as a result of the personalization, with average cosine similarity of 0.538, and is also able to ground its responses in the provided source materials (which reduces hallucination). These evaluations collectively show the efficacy of our personalization and guardrails to support LLMs in combating misinformed beliefs.

Our research shows that the generative AI can be enhanced with both personalization and external source materials, to craft diverse and grounded responses, with the specific goal of combatting misinformed beliefs. While most current approaches to countering misinformation largely overlook individual user characteristics [11], our methods emphasize tailoring information and aligns with work that suggests the effectiveness of targeted persuasion over non-targetted persuasion [10, 32]. The paper makes the following contributions:

- We introduce an LLM agent pipeline that is grounded in facts, and can help reduce an individual's misinformed beliefs through personalization.
- Our proposed pipeline is a mix of traditional ML methods and LLMs, ensuring that the final output is interpretable.

2 Background

This section explores prior work related to our research, providing theoretical groundings for selecting the three persuasion styles, and discusses existing research in the intersection of LLMs and persuasion. Moreover, we explore current work in RAG models, establishing the advantages of using a RAG model in the pipeline.

2.1 The Theory of Persuasion

In his manuscript "Rhetoric", Aristotle defined three kinds of persuasion that affect the audience: "personal character of the speaker", "putting the audience into a certain frame of mind", and, "proof, or apparent proof, provided by the words of the speech itself" [8]. Today, these are framed as ethos, pathos, and logos, and they are commonly used as measures of persuasion in the psychology and social sciences domain. Ethos is appeal to the credibility of the source of the information, pathos refers to the emotions evoked by the statement, and logos refers to the evidence/logic provided in a given argument [8]. In this research, we incorporate these three classical persuasive approaches; that is, we generate tailored messages that employ the approach that best fits the targeted audience member (based on their demographic characteristics and personality). This aligns with work that shows targeting is often more effective than non-targeted persuasion [21].

Extending Aristotle's initial theories of persuasion, models have appeared within psychology to characterize the effect of a message on an individual, emphasizing the role the individual plays in message reception. Persuasion, which can be thought of as an attitude change [19], can be pursued through two routes according to the Elaboration Likelihood Model [27]: central or peripheral. In the central route, opinions are formed around argument strength, while in the peripheral route, opinions are formed more on heuristics. These two methods of persuasion can be utilized in tandem to determine the effect of the message on the individual [19]. In turn, persuasion is highly reliant on the individual receiving the message, and personalization can play a large role in successfully expressing a message effectively [34, 44].

2.2 LLMs and Persuasion

As use of LLMs has become more ubiquitous, researchers have looked into the intersection of persuasion and LLMs. While persuasion can be used for malicious purposes such as convincing people to believe falsehoods [7], there is potential for LLMs to utilize persuasion for the positive goal of countering misinformation [18]. A recent study exploring the persuasiveness of LLM models suggested that simply prompt-engineering a base Llama-2-70B-chat model can "generate persuasive arguments that incorporate dimensions of

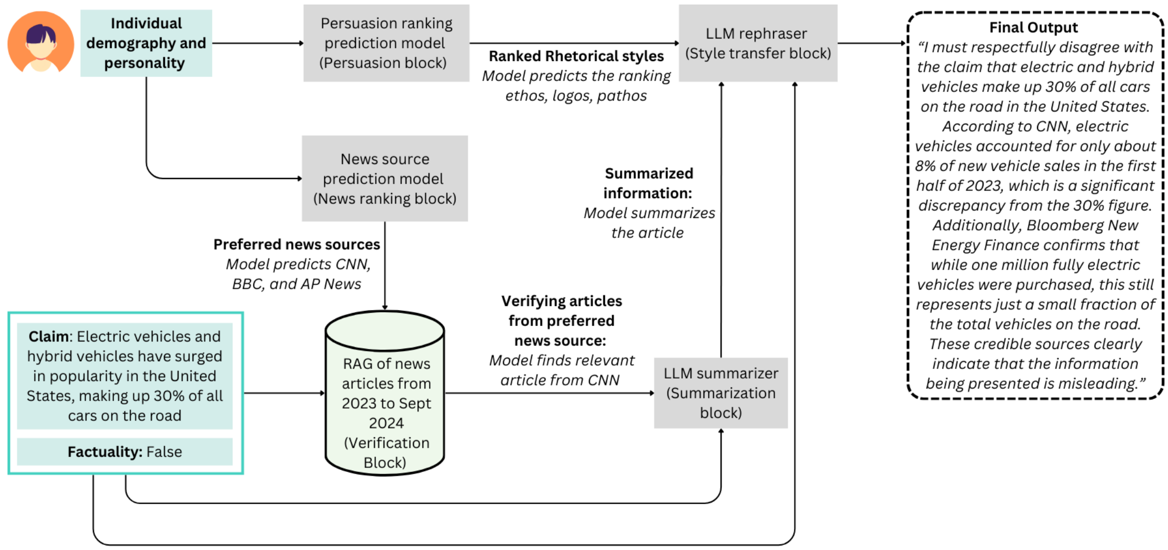


Figure 2: Our proposed pipeline aimed at providing personalized information to help users reduce belief in political misinformation. Individual demography and personality is used to predict preferred news source and rhetorical styles. With the claim and preferred news sources as input, RAG is queried to find verifying articles from preferred news sources. These articles are summarized, and then rephrased to incorporate the preferred rhetorical style. This rephrased information is the final output.

social pragmatics underpinning established psycho-linguistic theories of opinion change" and that these arguments had persuasive impacts on individuals [4].

Another study focused on measuring the persuasive power of LLM-generated responses based on personality traits of individuals [21]. Commonly used to measure personality, the "Big Five" personality traits are a means to describe an individual through openness, conscientiousness, extroversion, agreeableness, and neuroticism [23], and "Big Five" personality traits have been previously shown to affect user engagement [14]. For each of these features, there is a range that an individual can fall into. For example, an individual on the extroversion scale could be more introverted, or extroverted and social. In the study by Matz et al., ChatGPT-3 was prompted to write persuasive statements to cater to opposite ends of each of these features for domains such as marketing, political appeals for climate actions, and so on. Individuals then selected the more persuasive statement, and they were able to find openness, conscientiousness, and extroversion played a role in determining what an individual would prefer; agreeableness was found to have no effect and neuroticism was not tested. This study highlighted how ChatGPT can create tailored messages based on personality traits, and that individuals are receptive to this persuasion method [21].

2.3 Retrieval Augmented Generation

Retrieval Augmented Generation [17], or RAGs, provide a way to empower LLMs with specific data without prior fine-tuning. In the original study where RAG was introduced, the RAG method was tested on question-answering task on open domains, abstracts, Jeopardy, and fact verification [17]. Their study found that the RAG system achieved state-of-the-art performance, offering key advantages including effective document marginalization, generation of accurate answers, and reduced hallucination in responses [17].

Since the original study, RAG models have been used extensively in combination with LLMs across various domains, especially when factual accuracy is important, such as in misinformation and fake news detection. Khaliq et al. explored using RAG for multimodal political fact-checking with textual claim and a corresponding image as context [15]. They also explored enhancing the reasoning capabilities of RAG using Chain of RAG (CoRAG) and Tree of RAG (TreeRAG) [15]. Similarly, Markey et al. utilized a RAG for drafting medical documents from LLM responses [20]. The incorporation of the RAG was found to improve writing quality and accuracy of up-to-date information [20]. These examples suggest that integrating RAG can improve performance of LLM response generation in information niches, thus making it suitable for political domain.

3 Methods

We provide an overview of the proposed agent pipeline (Figure 2). The agent consists of multiple components - 1) News-source ranking block (subsection 3.3), which predicts potential news sources that the user might like, 2) Verification block (subsection 3.4), which pulls in information from a RAG model to verify the claim, 3) Summarizing block (subsection 3.5), which summarizes the information pulled from the RAG model in a few sentences, 4) Persuasion Rank block (subsection 3.6) to determine the persuasion style most important to the user, and 5) Style transfer block (subsection 3.7), which rephrases the summarized information to match the persuasion style. To build the news ranking and persuasion ranking prediction models, we collect and annotate data for training (subsection 3.1). Moreover, we also build a RAG model which is used in the verification block (subsection 3.2). The codebase for our pipeline can be found in this github link: https://github.com/ROC-HCI/Personalizing_LLM_Responses_Pipeline.

3.1 Data Collection and Annotation

To extract relationships between an individual's demographics and personalities with news trust and persuasion preferences, a dataset is built to train the news source ranking block and the persuasion ranking block. We first curate a series of political misinformation statements and their corresponding verifying information, and we also select a set of diverse news sources. These are then annotated by 338 individuals, with each individual annotating three samples.

3.1.1 Curating political misinformation statements and their corresponding pieces of verifying information. We start by collecting and curating a series of political misinformation statements and their corresponding pieces of verifying information. The misinformation statements are related to political parties or policies, and are statements with objective ground truth.

The data collection is done in three different ways to include variety. First, we conduct a series of in-person focus group studies, where participants craft their own misinformation, and find its corresponding verifying information online. Participants then discuss the believability of each of the misinformation statements they came up with, and through this process, we qualitatively validate that people's perceptions and emotional reactions to misinformation differ. Second, we look through online sources such as Politifact [13], AP News (specifically, a column called Not Real News) [36] and FactCheck.org [35], to find the most popular online political misinformation and its debunking evidence. Finally, researchers also generated 100 misinformation statements, and found verifying information from news sources such as groundnews.org [1], AP News, Reuters and so on.

All the misinformation statements and corresponding debunking information gathered using the above ways is then fact-checked by two annotators. The debunking information is stylistically modified to be unambiguous and directly address the claim. Overall, we generate 100 such pairs.

Next, seven different paraphrased versions are made of the verifying information using a GPT 3.5 model, where the model is prompted to come up with different paraphrased variations of the same verifying information. These paraphrased versions are reviewed and edited by two separate annotators to ensure that they all contain the same information. Through this process, each political misinformation claim has seven different paraphrased versions of the same verifying information.

3.1.2 Selection of news sources. A list of ten news sources is selected, ensuring a balanced representation of right, left, and center-leaning sources. We use Ad Fontes [2], All sides Media bias chart and the ratings in groundnews.org [1] to determine the political leaning of the news source. Our final news sources and their respective political leanings are Washington Examiner (right), New York Post (right), Fox News (right), AP News (center), The Economist (center), BBC (center), New York Times (left), CNN (left), Washington Post (left), and MSNBC (left).

3.1.3 Data annotation. After curating the political statements (including corresponding verifying information) and the set of news sources, we set up an experiment to understand individuals' preferred news sources and persuasion, recruiting participants through MTurk. Each annotator first rates on a 5-point Likert scale the believability of the misinformation statement (where the scale ranges from "Strongly Disagree" to "Strongly Agree"). Then, they are shown seven different verifying information with different news sources attached to them, and asked asked to rank the top three verifying information-news source pairs. Since the verifying information essentially contain the same information, the ranking choices made by participants reveal their news sources preferences. Participants also provide demography and personality-related (using a shortened version of the Big Five Questionnaire [29]) data and rank the relative importance of ethos, pathos, and logos to them when assessing information. This data is used to train the news ranking block (using individual persona and news ranking information), and the persuasion block (using individual persona and their persuasion rankings). 338 participants in total participated in this data collection, leading to a sample size of 1014 annotated data (as each participant contributed 3 data points).

3.2 Implementing Retrieval-Augmented Generation

We utilize Retrieval Augmented Generation (RAG) to provide the pipeline with accurate, up-to-date information. The RAG is hosted through ChromaDB, an online platform that allows for easy storage and retrieval of data, and populated with full-length news articles from the same 10 news sources as mentioned before. Utilizing GNews, an API to pull news headlines from a given news source over a given time period, and Newspaper3k [25], an API to scrape full articles based on an article URL, 100 full length articles were requested from each source for every week in 2023 and 2024 (until the first week of September). By only considering the most recent articles, we can ensure the content will be (1) most up to date and (2) most likely contain accurate information. After collecting all full length articles, articles that experienced errors during the retrieval process, articles containing the words "opinion" or "editorial", and duplicate articles were removed from the dataset. Overall, we curated 72594 full articles, having removed 1715 duplicates.

After compiling the final set of articles for the RAG, each article was encoded using the 'all-MiniLM-L6-v2' model, as it is native to ChromaDB. Finally, the list of full length articles and their embeddings were stored using ChromaDB to be used later for retrieval.

This database can then be queried, and the results pruned to only include news sources the individual trusts. The statement to be proven or disproven is provided to the RAG, and the Chroma query returns the top 10 articles related to the query. For each article returned, if the news source the article belongs to is included in the user's predicted list of trusted sources, that article will be considered. If none of the articles returned from the query are from a source in the individual's trusted source list, the top two relevant articles will be returned to ensure the pipeline has facts to base its response on in future steps.

Rhetorical Style	Generated Response
Ethos Response	I must respectfully disagree with the claim that electric and hybrid vehicles make up 30% of all cars on the road in the U.S. According to credible reporting from AP News, electric vehicle sales are projected to reach only 9% of all passenger vehicles this year, a significant difference from the 30% figure. Additionally, the U.S. is still behind other countries in EV market share, which further supports that we are not at that level of adoption. Lastly, the higher average cost of new electric vehicles indicates that affordability is still a significant barrier for many consumers, making widespread adoption at that level unlikely.
Pathos Response	I understand how easy it is to get caught up in the excitement surrounding electric vehicles, especially with all the buzz about their growing popularity. However, the reality is that electric vehicle sales in the U.S. are projected to only reach about 9% of all passenger vehicles this year, which is significantly lower than the 30% claim. It's important to recognize that while progress is being made, we still have a long way to go compared to other countries leading in EV adoption. I hope this clarifies the situation and helps us have a more accurate understanding of the current landscape.
Logos Response	The claim that electric and hybrid vehicles make up 30% of all cars on the road in the U.S. is misleading. According to AP News, EV sales are projected to only reach 9% of all passenger vehicles this year, which starkly contrasts with the 30% figure. Additionally, the U.S. still trails behind other countries in EV adoption, indicating that our market is far from saturated. The higher average cost of new EVs further suggests that many consumers are still opting for traditional vehicles, undermining the assertion of widespread EV popularity.

Table 1: Examples of final outputs, based on different rhetorical preferences, assuming results from all other blocks remain the same. Ethos emphasizes on the credibility of the source of the information, pathos refers to emotional appeal, and logos refers to the logic of the argument. Each response is generated using the same personality and false statement: "Electric vehicles and hybrid vehicles have surged in popularity in the United States, making up 30% of all cars on the road."

3.3 News Source Ranking Block

We use collaborative filtering, with K-means clustering to predict people's news source preferences. As features, we consider their personality and demographic information. For every new user, we consider the top five similar user and their preference in news sources. The weighted score for the news source preferences for the similar user, is used as the prediction for the new user.

3.4 Verification block

The verification block serves the purpose of identifying relevant articles for a given statement and selecting only the articles written by predicted trusted sources using RAG. The verification block takes in the ranked news sources (the output from the news source rank block), the statement, and the statement's factuality as input and provides a list of M sources with full length articles regarding the topic.

3.5 Summarization Block

The content of the selected new articles is summarized through prompt-engineering using GPT 4o-mini model. The number of tokens is set to 1000 to ensure that the output of this block captures all of the relevant information but is still somewhat concise. The summarization block takes in the full articles of the trusted sources (the output of the verification block), the statement to be proven or disproven, and the factuality of the statement as input. As output, the summarization block provides a concise summary of K relevant articles (where K <= M). To add appropriate guardrails to protect users and ensure that they do not encounter offensive language, the prompt ends with "Do not use any hate speech or vulgar language".

3.6 Persuasion Block

We use collaborative filtering, with K-means clustering to predict people's preferred ranking of the persuasion styles. Considering their personality and demographic information as features, we first find the top five similar users. Next, we consider their ranking of the ethos, pathos, and logos, and do a weighted sum. This provides the ranking of the persuasion styles for the user.

3.7 Style Transfer Block

The last block of the pipeline is the style transfer block, which also uses the GPT 4o-mini model. The style transfer block takes in the summarized articles (output from the summarizing block), the dominant style for the person (output from the persuasion rank block), the statement to prove or disprove, and the factuality of the statement as input and provides a presentation of the facts in the dominant style. This is the final and only output that the end user sees. Table 1 shows the different outputs based on different persuasion ranking.

We use the same setup as the summarization block except for the maximum amount of tokens, which is set to 150. This allows the response to be concise, clear, and simple since the output should be 4 sentences or less. The factuality of the statement to prove or disprove directly affects the prompt fed into the model. The prompt begins, "You are an informed citizen, persuading another citizen that the 'Information' is true / false".

For true statements, the prompt starts with, "Given that the 'Information' is true, rephrase the 'Summary' to prove the 'Information' by..." and is reversed for the false case. Since we are attempting to speak to an individual's preference on style, we want to either emphasize "the credibility of the news sources", "empathy towards the reader", or "the impact of the evidence", which represent ethos, pathos, and logos. This grounds our pipeline in the basic principles of persuasion, presented by Aristotle [8]. We also prompt the model to respond as an informed citizen, limited to four sentences, and remove any summaries irrelevant to the statement. Finally, to ensure that the model does not use offensive language, we bar the use of hate speech or vulgar language as part of the prompt.

4 Evaluation 1: Evaluation of Individual Blocks

In this section, we discuss our evaluation results at the block level. Evaluating each block separately allows us to identify their individual strengths and weaknesses, while also providing insight into the overall performance of the pipeline.

4.1 News Ranking Block Validation

To validate our news ranking block, we perform a stratified k-fold cross validation as our model aims to predict the top news sources for the individual. For evaluation, a prediction was deemed successful if the model correctly identified at least one of the individual's top news sources, assigning a score of 1 to that prediction. Predictions that did not meet this criterion were assigned a score of 0. The success rate for each fold was computed as the ratio of correct predictions to the total number of predictions in that fold. Considering k=10, the average success rate was determined by taking the mean of the success rates across all folds, giving a value of 0.82.

Prior literature suggests that most individuals exhibit moderate preferences for news sources, with only some highly politically-charged individuals watching partisan news [28]. This means that the difference in scores for each of the news sources might be very slight in most cases. This characteristic inherently increases the difficulty of the prediction task. Within the context of our work, the goal of this block is to maximize the chance that at least one of the news sources mentioned by the pipeline is preferred by the individual. Therefore, our evaluation here suffices for the intended purpose of the system.

4.2 Verification Block Validation

Validating the efficacy of the verification block response is challenging because there is no golden set of correct information to provide. A statement may be true, or untrue, for several reasons. For example, the statement, "Only Americans under 50 are affected by the increased recent inflation" is false for several reasons: (1) inflation simply impacts everyone who lives within the United States, (2) inflation can pose unique challenges for individuals who are older, for example, those who draw from social security [16], (3) the price of most items are affected by inflation, including items such as cable TV [31] that may skew towards older populations, and so on. Therefore, there is no "correct" response to debunk this statement, other than providing some information that discredits it, which as shown, could take on a variety of forms. Because personality and demographics are inputs into the system, we expect this output to look different based on those attributes, further complicating the evaluation.

TruLens, a software tool, offers a way to validate effectiveness between the query and the resulting articles, the retrieved articles and the summarization, and the summarization to the intended goal of the project [9]. The "RAG Triad" is aimed at quantifying context relevance, groundedness, and answer relevance of a RAG system. Context relevance investigates how the retrieved context (in our case, full articles) are relevant to the original query (asking to prove or disprove a given statement). Groundedness investigates how the LLM response (the summarization of the articles) is supported by the context. Answer relevance provides insight into how relevant the response is to the original query (ensuring the model did not hallucinate or stray from the original purpose). Each of these metrics is on a scale from 0 to 1, with 1 being the best score for all.

Statement	Factuality
The United States does not produce the most oil in the world.	False
Electric vehicles and hybrid vehicles have surged in popularity in the United States, making up 30% of all cars on the road.	False
While gas prices are expensive in certain states, it is consistently most expensive in California and West Coast states.	True
In 2023, only 10 states saw a population increase.	False
United States population trends have never recovered since COVID in 2020.	False
The provisional number of births for the United States in 2023 was 3,591,328, down 2% from 2022.	True
Only Americans under 50 are affected by the increased recent inflation.	False
Prices for used cars and trucks, and new vehicles have fallen since 2023.	True
The average price for airline tickets increased in August 2024 when compared to August 2023.	False
The number of people in the U.S. illegally is upwards of 20, 25, maybe 30 million.	False
In January 2023, the U.S. started accepting people monthly from Cuba, Haiti, Nicaragua and Venezuela under a humanitarian parole program.	True
New York City has been accommodating illegal immigrants in luxury hotels.	False

Table 2: Statements and associated factuality used in evaluations

To set up the RAG Triad evaluation, we utilized a TruLens custom wrapper function, with the backbone of our previously curated RAG and the three feedback measures (context relevance, answer relevance, groundedness). From there, we integrated our verification block and summarization block setup within the wrapper. The process queries our RAG to receive relevant source material and then generates the summarization as if it were a part of the pipeline. When running the process, TruLens monitors the inputs and outputs for each step and then calculates the feedback scores for the internal setup.

To analyze our verification and summarization blocks, a set of 12 statements and their factuality were provided as inputs into the system (Table 2). For each statement, the RAG was queried and a summary of the sources were generated. After completing the query and summarization process for all 12 statements, TruLens provided the process-wide feedback measures. Groundedness and answer

Algorithm 1: LLM Evaluator Workflow

$person \leftarrow 0$
while $person \leq 160$ **do**
  $statement \leftarrow 0$
  while $statement \leq 12$ **do**
    $output \leftarrow Pipeline(person, statement, factuality)$
    $output\_score \leftarrow$
      $LLMEvaluator(output, statement, factuality)$
  end
  $statement \leftarrow statement + 1$
  $person \leftarrow person + 1$
end

relevance will be discussed further in the summarization block evaluation. The context relevance score was 0.9104, suggesting that the articles provided in the RAG are robust and comprehensive.

4.3 Summarization Block Validation

The parts of the RAG Triad method relevant to the summarizing block are the answer relevance and the groundedness. This setup was fed the relevant articles from the RAG query, and then asked to summarize the information. With these parameters, the answer relevance was 0.9917 and the groundedness was 0.8884. These results support the notion that the summarizations tend to prove or disprove the statements provided, and the RAG can support summarizations through relevant articles.

4.4 Persuasion Block Validation

We run a k-fold cross-validation with k=10, and the model has to correctly predict the order of the three persuasion styles. Our accuracy is 0.208, where we consider a prediction to be correct if the models gets the entire ranking correct. Theoretically, the probability of predicting ranking of ethos, pathos, and logos correctly in order is 0.167, and hence our block out-performs the theoretical ranking.

5 Evaluation 2: Evaluation of the Final Output

In this section, we evaluate the final output of the pipeline. In order to judge both effectiveness and diversity within the pipeline, we conducted three different analyses on the final outputs of the system. The first utilizes an LLM judge to rate the effectiveness of the pipeline output, as done in prior work [4]. Second, we analyze the uniqueness of responses across the specific input statements provided. Finally, we analyze the robustness of the final output in accurately incorporating correct sources and not hallucinating.

5.1 Evaluation Using an LLM Judge

In order to complete this evaluation, 160 participants provided their personality and demographic data. For each personality, the response was generated for 12 unique statements using the pipeline methodology. After generating each of these responses, a separate GPT 4o-mini instance acts as an evaluator. The evaluator is given the ground truth of the statement and the statement itself, and rates each generated response individually on a discrete scale from 1 to 5, with 5 representing 'extremely convincing' and 1 representing 'not convincing'. Algorithm 1 represents the workflow for generating the responses and evaluations.

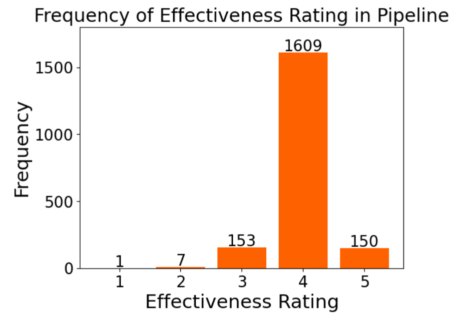


Figure 3: Frequency of the effectiveness rating by LLM evaluator. Given the pipeline-generated output, the LLM evaluator scored the output on a scale of 1-5, based on how convincing the output was at supporting or countering the original statement, with 5 being the most convincing.

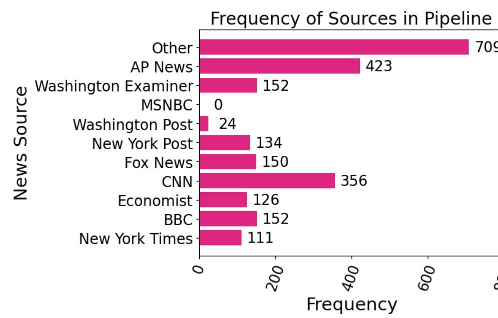


Figure 4: Frequency of source occurrence in the pipeline-generated output. Given all pipeline-generated outputs, the graph shows the total instance of each source.

Results from the 1,920 generations and comparisons found that the evaluator tended to rank the pipeline responses highly, with an average score of 3.99 across all responses. Figure 3 shows the frequency of the scores given by the evaluator.

Looking at the responses provided by the pipeline, we see a diverse representation of news sources, shown in Figure 4. Notably, there seems to be some preference in news source, for AP News and CNN. However, there is a lack of preference for MSNBC. While the distribution of the 10 identifiable sources deviates from a uniform distribution (Chi-Square test statistic = 958.622, p = 1.438e-200), we emphasize representation in most sources. We also see a large category of 'Other', which encompasses information from outside of the 10 expected news sources. This is further investigated in the hallucination detection subsection (section 5.3).

5.2 Cosine Similarity of Responses

Because the pipeline utilizes personality to curate the response, it is important that the responses between individuals are somewhat unique. To understand this, we conducted cosine similarity between responses, eliminating evaluations against a response generated for the same person. For each prompt statement, all outputs were compared against each other, resulting in 25440 comparisons per input. Figure 5 shows the cosine similarity between all outputs

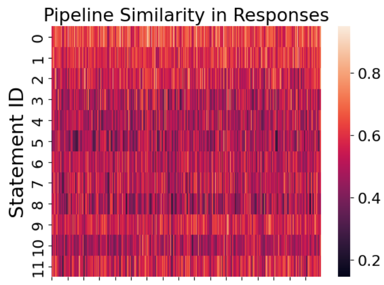


Figure 5: Cosine similarity of outputs based on statement, removing comparisons between outputs for the same person. Comparisons are made across all outputs for a given statement, as they would be expected to have the most similar responses. Lower scores indicate more diverse responses.

Statement ID	**0**	**1**	**2**	**3**	**4**	**5**
Average Cosine Similarity	0.654	0.603	0.564	0.504	0.494	0.477
Statement ID	**6**	**7**	**8**	**9**	**10**	**11**
Average Cosine Similarity	0.520	0.528	0.487	0.567	0.497	0.561

Table 3: Average Cosine Similarity of Outputs Based on Statement, Removing Comparisons Between Outputs for the Same Person

for a given statement. Table 3 shows the average cosine similarity score per input statement. We find the average cosine similarity across all statements to be 0.538, indicating that the pipeline does produce diverse outputs when given different personalities and demographics as input.

5.3 Hallucination Detection

Given that we know the pipeline is being provided with real, external sources, the hallucination efforts focus on hallucination regarding misattributions. In Figure 4, we note a large amount of "Other" citations, or citations that fall outside of the expected 10 sources. It is important to determine whether this is hallucination (where the "Other" source did not show up at all, or there is no source attribution) or rather an internal attribution (where, for example, CNN cites Pew Research Center, so the final output cites Pew Research Center). Given the articles pulled for generation, we aim to verify whether the sources mentioned (1) come from our list of 10, (2) are referenced by our list of 10, or (3) are not found anywhere. The tiers for this evaluation is listed below:

- Tier 1: (source) referenced is one of the set of 10 and returned in RAG query
- Tier 2: (source) referenced is not one of the set of 10; another source referenced by set of 10
- Tier 3: (source) referenced is not one of the set of 10; source is completely fabricated and not mentioned by any articles used for generation
- Tier 4: no (source) provided

As the presentation of the sources in the final output and the original articles can be somewhat inconsistent, potential sources were identified by looking for strings of words beginning in capital letters (as this seemed to be the only consistency). These words or strings of words were pulled out and then searched for within the original documents. When words or strings of words were not found, this was deemed a hallucination. Of course, this method has the potential to pull non-source words, such as words at the beginning of sentences or proper nouns such as "United States" or "Saudi Arabia". To try and correct this, a check was then run to search for each individual word within a string of words to verify whether it occurred in the original document. As the LLM often includes filler words (such as "Additionally") and abbreviations (such as "CDC"), we consider these estimates to be the upper bound of pipeline misattributions. Both versions are reported below.

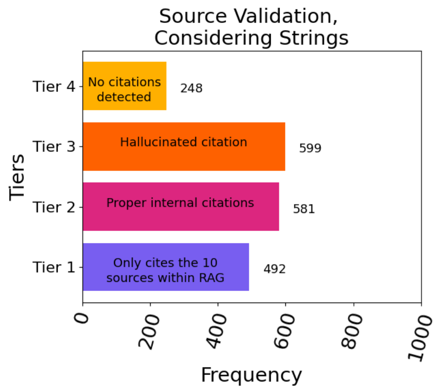
Of the 1920 final outputs generated by the pipeline, 248 had no detectable sources cited. Figure 6a shows the results for the full string search within the pulled sources. While the majority of citations fall within tiers 1 and 2 (acceptable for our pipeline), we see a large amount of sources within tier 3. After running a check on each individual word within a string (with the hopes of eliminating non-sources that were detected as such), we find a more acceptable minority of sources fall within tier 3. Figure 6b reports the results of the updated search and categorization.

6 Discussion and Limitations

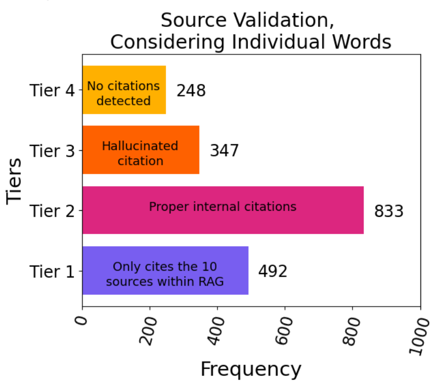
The potential for dual-use of LLMs as a tool for both mitigating and amplifying misinformation means that it is critical to study their impact, and ensure appropriate guardrails are put in place for their use. We take several measures to better ground our agent in facts, thus reducing scope for hallucination. For example, as we provide the system with both the claim and the factuality of the claim, conflicting information that does not match the factuality label would not be considered. Our findings suggest that LLM agents can be designed to be grounded in facts, and therefore, they can better counter misinformed beliefs.

Moreover, while one might argue that persuasion tactics can have malicious intent, it must be noted that our goal is to only tackle misinformed beliefs. On using personalized systems like this, it must be only deployed and used in spaces where people are open to such an intervention and they have given their consent. To build our system, participants agreed to share their demographic and personality data, along with their responses to the tasks. Similarly, future work is needed to understand what data individuals are willing to share for personalization tasks in real-world settings.

One notable strength of our pipeline for both news ranking and persuasion prediction lies in the use of collaborative filtering. While this approach is relatively traditional, it offers two key advantages. First, it ensures a level of interpretability in the predictions, making the model's outputs easier to understand and trust. Second, by employing traditional machine learning algorithms that can be trained locally, our approach addresses privacy and security concerns by avoiding the need to share sensitive data with LLMs. That said, it is worth acknowledging that methods for locally training LLMs are emerging, and in the future, might be worth looking into.



(a) Source attributions based on capitalized strings.



(b) Source attributions based on individual capitalized words.

Figure 6: Comparison of pipeline source attributions based on two different detection methods.

The goal of our study was to achieve reasonable performance in each of the blocks, ensuring the overall system functions effectively. Future studies can focus on improving the performance of each of the blocks and also evaluating them across a variety of metrics. This includes opting for larger datasets for persuasion and news source preferences, while also finding more interpretable evaluation metrics (unlike Trulens [9], for which some details of the model evaluation mechanism are hidden). Moreover, getting data from a large representative sample of participants to train the models would also help reduce potential biases in user modeling.

Due to resource constraints, we did not focus on directly comparing our pipeline with baseline models. However, standard LLMs are subjected to a training cutoff date (with information only available till the date it was trained), and the data used is also proprietary (and unknown to the end user) [3]. Our pipeline is not limited by training cutoffs of standard LLMs as it draws information from more recent news sources and our source bank is well-documented, which highlight the potential benefits of our system. Moreover, while we used GPT models for our pipeline, it might be worth exploring how the pipeline would perform if the model was replaced with other foundation models such as Gemini [33], Llama [37], and so on. Testing on multiple models would help improve the generalizability of our work.

Finally, large-scale user studies are needed to evaluate how the pipeline performs in a real-world setting. This includes understanding whether people are open to such an intervention and analyzing the impact of such an intervention both short-term and long-term, especially on social networks. Future research can involve designing various empirical studies to evaluate how participants interact with the pipeline, and also with each other after interacting with the pipeline.

7 Ethical Considerations

The authors of this paper would like to recognize the ethical implications of the proposed work. First, in order to deploy such a system on a large scale, it is imperative to receive consent from the users of that system to receive personalized content. We also recognize that technology can often be utilized for both beneficial and malicious use, especially persuasive methods. Extensive research has explored both the positives and the negatives associated with personalized persuasion [22, 42, 43]. Thus, it is crucial to build systems with safeguards, such as our design which incorporates credible articles as the backbone. Finally, we attempt to mitigate use of our system by bad actors by withholding the specific data used for building our model, while keeping the code open-source to promote research in this field. Any researcher interested in using our data can request access. We emphasize that all future researchers utilizing our system should gather large quantities of data and rely on various credible sources of information to reduce bias.

8 Conclusion

Our work demonstrates the ability to incorporate personalization into LLMs for positive social change: combating misinformed beliefs. We propose and evaluate a RAG-based LLM pipeline that personalizes information by considering user demographics, personalities, and trust in credible sources. Our evaluation of the system suggests that this approach improves both the diversity of the messages, while also reducing hallucinations and improving groundedness. The pipeline generates convincing responses, as indicated by its average rating of 3.99 out of 5 for response persuasiveness. Our findings contribute to the broader research goal of understanding personalization using LLMs.

Acknowledgments

We would like to thank Goergen Institute for Data Science and Artificial Intelligence (GIDS AI) for funding our work. We also acknowledge Rusmiya Aqid for her contributions to the project, and thank all the participants who joined our in-person focus group studies.

References

[1] [n. d.]. Ground News — ground.news. https://ground.news/. [Accessed 08-11-2024].
[2] [n. d.]. Research Guides: News Media Across the Political Spectrum: Starting Point: 1. "The Chart" — guides.library.harvard.edu. https://guides.library.harvard.edu/newsleans/thechart. [Accessed 08-11-2024].
[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal

Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
[4] Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 152–163.
[5] Yupeng Cao, Aishwarya Muralidharan Nair, Elyon Eyimife, Nastaran Jamalipour Soofi, KP Subbalakshmi, John R Wullert II, Chumki Basu, and David Shallcross. 2024. Can Large Language Models Detect Misinformation in Scientific News Reporting? *arXiv preprint arXiv:2402.14268* (2024).
[6] Fidelia Cascini, Ana Pantovic, Yazan Al-Ajlouni, Giovanna Failla, and Walter Ricciardi. 2021. Attitudes, Acceptance and Hesitancy among the General Population Worldwide to Receive the COVID-19 Vaccines and their Contributing Factors: A Systematic Review. *EClinicalMedicine* 40 (2021).
[7] Sijing Chen, Lu Xiao, and Jin Mao. 2021. Persuasion strategies of misinformation-containing posts in the social media. *Information Processing & Management* 58, 5 (2021), 102665.
[8] Edward Meredith Cope et al. 1877. *The rhetoric of Aristotle*. Vol. 2. University Press.
[9] Anupam Datta, Matt Fredrikson, Klas Leino, Kaiji Lu, Shayak Sen, Ricardo Shih, and Zifan Wang. 2022. Exploring Conceptual Soundness with TruLens. In *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 302–307.
[10] James N Druckman. 2022. A framework for the study of persuasion. *Annual Review of Political Science* 25, 1 (2022), 65–88.
[11] Lisa Fazio, David Rand, Stephan Lewandowsky, Mark Susmann, Adam J Berinsky, Andrew Guess, Panayiota Kendeou, Benjamin Lyons, Joanne Miller, Eryn Newman, et al. 2024. Combating misinformation: a megastudy of nine interventions designed to reduce the sharing of and belief in false and misleading headlines. (2024).
[12] Daniel J Flynn, Brendan Nyhan, and Jason Reifler. 2017. The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology* 38 (2017), 127–150.
[13] Poynter Institute. [n. d.]. PolitiFact. https://www.politifact.com/. [Accessed April 1, 2025].
[14] Nima Jamalian, Marios Constantinides, Sagar Joglekar, Xueni Pan, and Daniele Quercia. 2023. Our Nudges, Our Selves: Tailoring Mobile User Engagement Using Personality. In *IFIP Conference on Human-Computer Interaction*. Springer, 33–45.
[15] M Abdul Khaliq, P Chang, M Ma, Bernhard Pflugfelder, and F Miletić. 2024. RAGAR, Your Falsehood RADAR: RAG-Augmented Reasoning for Political Fact-Checking using Multimodal Large Language Models. *arXiv preprint arXiv:2404.12065* (2024).
[16] Lorie Konish. 2024. High inflation harms older households - and two factors determine who is most at risk, research finds. https://www.cnbc.com/2024/05/15/what-determines-how-well-older-households-cope-with-high-inflation.html
[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
[18] Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14279–14305. doi:10.18653/v1/2023.emnlp-main.883
[19] Andrew Luttrell. 2018. *Dual Process Models of Persuasion*. Oxford University Press. doi:10.1093/acrefore/9780190236557.013.319
[20] Nigel Markey, Ilyass El-Mansouri, Gaetan Rensonnet, Casper van Langen, and Christoph Meier. 2024. From RAGs to riches: Using large language models to write documents for clinical trials. *arXiv preprint arXiv:2402.16406* (2024).
[21] SC Matz, JD Teeny, Sumer S Vaid, H Peters, GM Harari, and M Cerf. 2024. The potential of generative AI for personalized persuasion at scale. *Scientific Reports* 14, 1 (2024), 4692.
[22] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences* 114, 48 (Nov. 2017), 12714–12719. doi:10.1073/pnas.1710966114
[23] Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60, 2 (1992), 175–215.
[24] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.
[25] Lucas Ou-Yang. 2013. Newspaper3k: Article scraping & curation. *Newspaper3k: Article Scraping & Curation-Newspaper 0.0. 2 Documentation* (2013).
[26] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744* (2023).
[27] Richard E. Petty. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change* (1st ed ed.). Springer New York, New York.
[28] Markus Prior. 2013. Media and political polarization. *Annual review of political science* 16, 1 (2013), 101–127.
[29] Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality* 41, 1 (2007), 203–212.
[30] Marten Scheffer, Denny Borsboom, Sander Nieuwenhuis, and Frances Westley. 2022. Belief traps: Tackling the inertia of harmful beliefs. *Proceedings of the National Academy of Sciences* 119, 32 (2022), e2203149119.
[31] Anita Snow. 2022. Working at 76: Inflation forces hard choice for older adults. https://apnews.com/article/inflation-older-americans-72c8a3ed29eb8b8183400bc161c0c9ee
[32] Ben M Tappin, Chloe Wittenberg, Luke B Hewitt, Adam J Berinsky, and David G Rand. 2023. Quantifying the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences* 120, 25 (2023), e2216261120.
[33] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
[34] Jacob D. Teeny, Joseph J. Siev, Pablo Briñol, and Richard E. Petty. 2021. A Review and Conceptual Framework for Understanding Personalized Matching Effects in Persuasion. *Journal of Consumer Psychology* 31, 2 (April 2021), 382–414. doi:10.1002/jcpy.1198
[35] The Annenberg Public Policy Center. [n. d.]. FactCheck.org. https://www.factcheck.org/. [Accessed April 1, 2025].
[36] The Associated Press. [n. d.]. AP News. https://apnews.com/. [Accessed April 1, 2025].
[37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[38] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The Spread of True and False News Online. *Science* 359, 6380 (2018), 1146–1151.
[39] Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426* (2024).
[40] 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6288–6304. doi:10.18653/v1/2023.findings-emnlp.416
[41] Longzheng Wang, Xiaohan Xu, Lei Zhang, Jiarui Lu, Yongxiu Xu, Hongbo Xu, Minghao Tang, and Chuang Zhang. 2024. Mmidr: Teaching large language model to interpret multimodal misinformation via knowledge distillation. *arXiv preprint arXiv:2403.14171* (2024).
[42] Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5635–5649. doi:10.18653/v1/P19-1566
[43] Brahim Zarouali, Tom Dobber, Guy De Pauw, and Claes De Vreese. 2022. Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media. *Communication Research* 49, 8 (Dec. 2022), 1066–1091. doi:10.1177/0093650220961965
[44] René Ziegler, Michael Diehl, and Anja Ruther. 2002. Multiple Source Characteristics and Persuasion: Source Inconsistency as a Determinant of Message Scrutiny. *Personality and Social Psychology Bulletin* 28, 4 (April 2002), 496–508. doi:10.1177/0146167202287007