MisinfoEval: Generative AI in the Era of "Alternative Facts"

Saadia Gabriel[♥] **Liang Lyu**[♠] **James Siderius**[♦]
Marzyeh Ghassemi[♠] **Jacob Andreas**[♠] **Asu Ozdaglar**[♠]
[♥]University of California, Los Angeles
[♠]Massachusetts Institute of Technology
[♦]Dartmouth College, Tuck School of Business

Abstract

The spread of misinformation on social media platforms threatens democratic processes, contributes to massive economic losses, and endangers public health. Many efforts to address misinformation focus on a knowledge deficit model and propose interventions for improving users' critical thinking through access to facts. Such efforts are often hampered by challenges with scalability, and by platform users' personal biases. The emergence of generative AI presents promising opportunities for countering misinformation at scale across ideological barriers.

In this paper, we introduce a framework (MisinfoEval) for generating and comprehensively evaluating large language model (LLM) based misinformation interventions. We present (1) an experiment with a simulated social media environment to measure effectiveness of misinformation interventions, and (2) a second experiment with personalized explanations tailored to the demographics and beliefs of users with the goal of countering misinformation by appealing to their pre-existing values. Our findings confirm that LLM-based interventions are highly effective at correcting user behavior (improving overall user accuracy at reliability labeling by up to 41.72%). Furthermore, we find that users favor more personalized interventions when making decisions about news reliability and users shown personalized interventions have significantly higher accuracy at identifying misinformation.

1 Introduction

In the last decade, there has been growing concern about the proliferation of misinformation on social media platforms. For example, between 2006 and 2017, sensational "fake news" articles spread rapidly on Facebook, diffusing farther and faster than truthful or reputable content (Vosoughi et al., 2018). These sharing trends have been amplified by "filter bubble" algorithms that intentionally create ideological echo chambers, which reinforce existing viewpoints and further facilitate spread of misinformation (Levy, 2021; Acemoglu et al., 2023).

Many proposed interventions for combating misinformation focus on tagging unreliable content (Clayton et al., 2020; Pennycook et al., 2019) or encouraging critical thinking by users (Lutzke et al., 2019; Pennycook et al., 2021b). However, the two major bottlenecks in many such interventions are user bias and scalability. Conventional fact-checking interventions rely on the assumption that users are rational agents, who will agree upon a common "ground truth" once exposed to enough information. This assumption is often violated in the real world due to the fact users do not process new information neutrally, and are more critical of counter-partisan news and more accepting of pro-partisan news at face value (Lord et al., 1979; Nickerson, 1998; Tappin et al., 2019). Tagging unreliable or suspicious content also requires careful inspection, which is currently performed by professional fact-checking organizations such as Snopes. These organizations are constrained both in terms of their financial resources and qualified fact-checkers they can employ. Alternatively, platforms have explored a decentralized approach through crowdsourced verification by users. While this approach allows for scale, it is vulnerable to user misuse such as recent reports of disinformation and partisan bias spread through X's Community Notes (Elliott and Gilbert, 2023).

Breakneck advances in large language models (LLMs) offer a promising avenue for large-scale fact checking, as they provide tools for fast information processing and can detect patterns associated with misleading content (Chen and Shu, 2023). Early evidence suggests that LLM-based explanations of veracity can significantly reduce social media users' tendency to accept false claims (Hsu

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 8566–8578
November 12-16, 2024 ©2024 Association for Computational Linguistics

et al., 2023). LLMs also present a potential path to understanding and countering user bias: recent work (Andreas, 2022; McIlroy-Young et al., 2022; Gabriel et al., 2022) argues LLMs are capable of very simple forms of world and cognitive modeling. This opens up the possibility of tailored approaches to countering misinformation that target users across diverse backgrounds (e.g. varying education levels or ideological leanings). Our agenda is to develop a powerful, automated tool via LLMs for generating tailored misinformation interventions and a framework for testing the effectiveness of these interventions (MisinfoEval). To do this, we have two study phases focused on examining effects of AI interventions with and without personalization:

Phase I: We conduct an A/B testing experiment using a simulated social media environment that features both true and false content.

Implications for Content Moderation: First we expand upon existing research showing that explanation-based credibility indicators can mitigate spread of misinformation. In a large-scale experiment, we show **explanation interventions can effectively inform diverse users about unreliable content, improving user accuracy over label-only indicators by at least 34.2% vs. 24.2% for label-only indicators**. Explanations generated by GPT-4 (OpenAI, 2023) perform best at encouraging user flagging of misinformation (pre-intervention users correctly flag in 3.1% of cases vs. 38.1% post-intervention). This indicates the promise of LLM-based explanations for future intervention strategies, corroborating findings from recent and concurrent work (Gabriel et al., 2022; Hsu et al., 2023).

Implications for Healthcare: The spread of medical misinformation poses a serious obstacle to healthcare providers if unchecked. Some notable recent examples include false claims spread online through websites like X (formerly Twitter), YouTube and WebMD of “alternative cures" for Cancer (Swire-Thompson and Lazer, 2019). Such false claims could pose a significant health risk given that alternative medicines can more than double the risk of mortality in Cancer patients (Johnson et al., 2017). **We find that GPT-4 explanations could provide automated support in mitigating medical misinformation, leading to an user accuracy at predicting news reliability of 97.6% (95% CI: [96.0, 99.2]).**

Phase II: Next, we explore how personalization of explanations can further improve their effectiveness. We measure how the degree of personalization affects user-reported helpfulness, and find that **explanations that are highly aligned with users' attributes (e.g., education, political ideology, gender) are deemed more helpful by users than explanations without personalization (average helpfulness score of 2.98 vs. 2.71)**. We also find that **users shown personalized explanations have significantly higher accuracy at identifying misinformation than users shown misaligned explanations**.

Our results highlight the powerful persuasive capabilities of LLMs. Our simulated social media feed platform and surveys with over 4,000 diverse participants will be made publicly available to further research on how LLM outputs influence users, including potential disinformation risks. To the best of our knowledge, this is the first study to consider tailored LLM-based interventions based on specific attributes of social media users. We view it as a first step in this agenda, since general advances in foundation models will increase these tools' capabilities for combating misinformation. We envisage better fine-tuning of explanations as additional information becomes available about users (without violating their privacy).

2 Related Work

Our work aligns with a growing body of literature, mostly from political science, that examines the effectiveness of mitigating misinformation through user-facing interventions. Assuming a known “ground truth” label determined by human or AI fact-checkers, this literature aims to reduce user consumption and interaction with false content by designing effective ways to present this information or to otherwise nudge users to consider them.

Fact-checking labels that are specifically attributed to AI have been shown to be effective as interventions in reducing user consumption of misinformation (Kyza et al., 2021), though earlier studies found they are often less effective than labels attributed to other sources such as professional fact-checkers (Seo et al., 2019; Yaqub et al., 2020; Liu, 2021; Zhang et al., 2021). There is evidence that explaining the mechanics behind how the

Figure 1: Examples of a post in the simulated newsfeed (left), and a pop-up intervention with a veracity label (right).

fact-checking label is generated improves their effectiveness (Epstein et al., 2021). Concurrently with our work, it has been found that GPT-based explanations of content veracity can significantly reduce social media users' reported tendency to accept false claims (Hsu et al., 2023), though they can be equally effective when used with malicious intent to generate deceptive explanations (Danry et al., 2022). There have also been some early works that explore the use of personalization in AI fact-checking systems, such as Jahanbakhsh et al. (2023), which examines the effects of a personalized AI prediction tool based on the user's own assessments; and Jhaver et al. (2023), with a focus on toxicity in personalized content moderation tools. Our work departs from these as we consider the generation of arguments and justifications given a label, rather than predicting the veracity.

3 Social Media Platform Experiment and Study Design

We recruit human participants to interact with a simulated news feed interface that mimics real-world social media platforms such as Facebook and X. The news feed consists of news headlines (claims), and an intervention button ("Find out more") that the user may voluntarily click on. Upon doing so, they may see an intervention with a veracity label of the headline (true or false) and an explanation of the label (see Figure 1). Users may react to the news item as they normally would on social media, and provide feedback on the intervention. By varying the types of interventions presented to users and comparing their subsequent behavior, we can analyze the impacts of these interventions. Details of the interface are given in §3.1. Phase I (§4.1) compares the effectiveness of five non-personalized explanations, while Phase II (§4.2) directly compares GPT-4 generated explanations with and without personalization. In both phases, the experimental interface consists of four components: (1) a consent form; (2) task instructions; (3) a questionnaire on user demographics and opinions; (4) a simulated newsfeed.

3.1 MisinfoEval Environment

Each participant receives 5 news items, randomly sampled from a dataset of 460 news headlines collected by (Pennycook et al., 2021a). Our experiment uses 188 true articles and 185 false articles. Each news item consists of the *headline* (which we also call a *claim*), the accompanying image and the source of the news article.[1]

User Interface Design. Users can interact with the posts by liking, sharing or flagging them (Figure 2, left). Each user is instructed to perform at least one of these interactions for at least three out of five news items. Users also have the option to click on a "Find out more" button, which displays a pop-up that we call an *intervention*. Except for the control setting, the intervention consists of two pieces of information: a *label* indicating whether the claim is true or false and an *explanation* either supporting or refuting the claim based on the label. In the pop-up, users can rate the perceived helpfulness of this information on a 4-point Likert scale (very helpful, somewhat helpful, somewhat unhelpful, very unhelpful). They can also indicate whether they believe the claim is true, false or are

[1]The dataset also includes misleading headlines, but we omit these to have a binary true/false label.

Intervention Type	Description	Example
Label Only	A simple ground-truth label indicator	***This claim is true/false.***
Methodology (AI)	Following from Epstein et al. (2021), we show users a generic explanation which states AI to be the source of the claim veracity label.	***This claim was verified/refuted by an AI model trained on a large-scale corpus of web data.***
Methodology (Human)	Same as above, except the source is stated to be fact-checkers.	***This claim was verified/refuted by non-partisan fact-checkers.***
Reaction Frame Explanation	Following from Gabriel et al. (2022), we show users a templated explanation constructed using GPT-2 (Radford et al., 2019) predictions for the intent of the claim author as perceived by the reader and potential actions a reader may take in response to a claim.	***This claim is true/false. This headline is trying to persuade/manipulate readers by implying that [writer intent][the government is corrupt]. It is compelling readers to [reader action][want to find out more].***
Zero-shot GPT-4 Explanation	We instruct GPT-4 to generate an explanation for why a claim is false/true without user-specific information. The prompt given to GPT-4 is "***Write a short explanation for why the headline '[claim]' is '[label].' Do not mention that you are AI. The explanation must be less than 100 words.***"	***The headline is false because Dr. Anthony Fauci has not been arrested by any Special Forces. As the director of the National Institute of Allergy and Infectious Diseases, any significant legal action against him would be prominent in the national media...***
Personalized GPT-4 Explanation	We do the same as above, but with the following prompt that incorporates inferred user attributes ***"Write a short explanation for why the headline '[claim]***[Special Forces Arrest Deep State Dr. Anthony Fauci]***' is '[label]***[false]***' that will appeal to an [education]***[uneducated]***, [gender]***[male]***, [race]***[white]***, [age]***[18-29] ***year old reader with [politics]***[conservative] ***political beliefs. Do not mention that you are AI. Do not mention the type of reader. The explanation must be less than 100 words."***	***This headline is false because there's no evidence. Terms like 'Deep State' often get thrown around without proof. Remember, just because someone disagrees with Dr. Fauci's advice, doesn't mean he's part of a conspiracy. Don't believe everything you read on the internet, mate. Always consider the source.***

Table 1: Types of intervention methods used in this experiment.

uncertain.[2]

MisinfoEval Intervention Types. In Phase I, we consider 5 types of previously proposed interventions for misinformation mitigation. Each participant is randomly assigned to one of the five types of interventions with equal probability and only shown that intervention. In Phase II, we introduce a sixth intervention type: personalized GPT-4 interventions. Table 1 lists the six types of interventions, with examples based on the false claim "*Special Forces Arrest Deep State Dr. Anthony Fauci*".

In Phase II, we compare non-personalized and personalized GPT-4 interventions (the last two rows of Table 1). Personalized interventions tailor to a specific demographic group based on a set of attributes (gender, race, age, education level, political affiliation). Details on participant selection and attribute values used for the personalization study can be found in A.5.

3.2 Participants

In this section, we explain our methodology for user recruitment and qualification tasks we require users to undergo in order to ensure quality of results (e.g., filtering spammers).

3.2.1 Recruitment and Quality Control

We use the Amazon Mechanical Turk[3] crowdsourcing platform to recruit a diverse pool of 9,262 workers from the United States with at least a 98% HIT[4] approval rating as potential study participants. To filter spamming workers, we ask them two "attention checks" questions that require them to write out the minimum number of posts they must interact with (3) and the number of posts in the newsfeed (5). Any workers who fail either of these attention checks are disqualified from participating in the rest of the study. We also disqualify workers who fail to follow the instructions by interacting with less than 3 posts. Lastly, we filtered workers who completed over 10 HITs but always predict the same reliability label for news articles (either

[2]The exception is the control setting in Phase I, where the pop-up only asks users to evaluate the reliability of the claim.

[3]`https://www.mturk.com/`

[4]Human intelligence task

Intervention	Accuracy (% Correct)			False Content Sharing (%)		False Content Flagging (%)		Helpfulness (% Helpful or Very Helpful)
	Before	After	Δ	Before	After	Before	After	
Label Only	55.00 ± 2.66	79.25 ± 2.33	24.24	5.12	4.18	4.37	21.38	85.43
Reaction Frame	55.07 ± 2.09	95.84 ± 1.00	40.77	4.62	0.00	3.83	1.00	94.31
GPT-4 (non-personalized)	52.16 ± 3.60	93.88 ± 1.54	41.72	3.66	7.26	3.10	38.17	92.17
Methodology Explanation (AI)	51.87 ± 2.60	90.98 ± 1.35	39.11	9.59	16.95	2.82	28.95	88.66
Methodology Explanation (Human)	55.77 ± 2.04	89.97 ± 1.45	34.20	2.36	5.60	5.34	4.23	90.14

Table 2: Accuracy at ground-truth label prediction, changes in interactions and perceived helpfulness results for all intervention types, both before interventions (left column) and after interventions (right column). Accuracy is shown with 95% bootstrapped confidence intervals.

true or false). 4,950 workers passed the qualification tests. Volunteered information about worker demographics is given in A.6.

3.2.2 Selection of User Attributes for Personalization

Prior work (Santurkar et al., 2023) has shown that identity groups based on the attributes we use for personalization (e.g. gender and political affiliation) have divergent beliefs around key social and political issues, which may affect their perception of news reliability. While many of these attributes are indirectly associated with beliefs, we also personalize based on political ideology, which is directly associated with beliefs. We leave exploration of how personalization influences specific beliefs (e.g. vaccine hesitancy) to future work.

4 MisinfoEval Environment Results

In 4.1, we discuss best practices for evaluating effectiveness of misinformation interventions. We first compare non-personalized interventions in a simulated social media newsfeed in 4.2. Then we assess the effectiveness of personalized interventions in 4.3.

4.1 Quantifying Effectiveness of Explanations

Progress in development of misinformation interventions has been hindered by a lack of standardization in evaluation. In line with the framework posited by Guay et al. (2023), we measure user interaction with *both* true and false content. Our aim is to capture *discernment* - the extent to which an user believes or intends to share false content relative to true content. We quantify effectiveness of interventions using two metrics:

- The objective effect of interventions on users' accuracy in recognizing misinformation and factual content.
- Users' perception of the interventions' helpfulness in performing misinformation detection.

While we also provide results for user interactions with claims, we emphasize accuracy and helpfulness as effectiveness metric given that interactions like sharing are not clear indicators of a user's belief in a claim (Pennycook et al., 2021b).

4.2 Phase I: Non-personalized Explanations

We measure the effectiveness of all five (non-personalized) interventions in mitigating misinformation, by comparing users' pre- and post-intervention behavior. We use the following metrics for the comparison: (1) accuracy of users' veracity prediction for each headline; (2) interaction with false headlines, such as sharing and flagging; (3) user-reported helpfulness of the label and explanation. Users are first shown the news without access to interventions and their accuracy is measured. We then show them the interventions and collect accuracy information again. If they do not view an intervention, no additional data is collected.

Comparison with Baselines: Table 2 shows results for all non-personalized intervention variations averaged over 3 randomized trials. For each intervention, we considered at least 604 instances of user-claim interactions (see Table 4 in A.4 for full statistics of interaction instances).

We find that without interventions, users consistently struggle to identify the true accuracy of news. All intervention types significantly improve users' overall accuracy (up to 41.72%). Also, all explanation-based interventions have a greater effect on accuracy than label-only interventions.

Interestingly, we find that effects on interaction behavior vary considerably across tested intervention types. In particular, interventions can actually *increase* sharing of false news. One hypothesis for this may be users wanting to fact-check claims with others they trust, since we also see an increase in false content flagging for 3 out of 5 interventions. In particular, Label Only, GPT-4 and AI methodology interventions led to relatively high rates (21.38-38.17%) of false content flagging by users. We hypothesize that the self-contained fact-check in the GPT-4 explanation may reduce users' interest in reaching out to trusted networks, though we do see an increase in false content sharing along with increased accuracy and flagging.

Overall, three interventions seem the most effective: Reaction Frame explanations, non-personalized GPT-4 explanations, and surprisingly, a simple methodological explanation that states the veracity label was generated using an AI model, without mentioning specifics of the claim. Reaction Frame and GPT-4 explanations are comparable in terms of accuracy improvement and similar in terms of self-reported helpfulness scores, though GPT-4 explanations consistently encourage more user flagging.

Analysis of Healthcare Misinformation: We further breakdown the results to assess domain-specific considerations in developing effective intervention mechanisms against medical misinformation. When we consider a balanced subset of 54 articles focused on medical real news and misinformation, we find that the prediction accuracy indicates there is no significant difference between trust in AI as a source (91.98% acc, 95% CI: [89.88, 94.07]) compared to a human fact-checker (92.33% acc, 95% CI: [90.20,94.47]). GPT-4 explanations are particularly promising interventions in the healthcare domain (97.65% acc, 95% CI: [96.03,99.27]), where there may be less disagreement due to partisan bias.

4.3 Phase II: Personalized Explanations

Our findings in the previous section indicate AI interventions can be effective at encouraging more informed, responsible behavior from users. However, this does not directly tackle risks of polarization and user bias. The next research question we seek to answer is whether GPT-4 interventions can be further improved through personalization designed to counter partisan behavior.

4.4 Self-reported Helpfulness Results

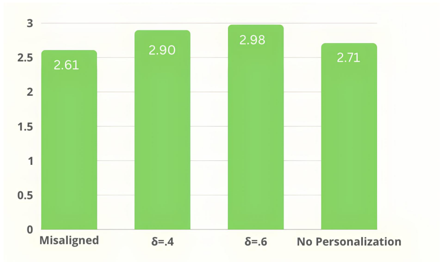
We measure the effectiveness of personalization by comparing user-reported helpfulness scores for personalized explanations and scores for non-personalized GPT-4 explanations. Figure 2a shows mean helpfulness scores based on 6520 observations of GPT-4 explanations without personalization and 3000 observations with personalization. We use a 0-1 score for degree of personalization alignment, based on how many of the user's ground-truth attribute values were used to generate the explanation (e.g a score of 0.4 means 2 out of 5 attributes used to generate the explanation match the user). We consider an explanation *aligned* with the user if its degree of personalization is at least = 0.4, and *misaligned* otherwise. We find 49.5% of the explanations are aligned, and the maximum alignment score is 0.6.

Overall, users find explanations of veracity more helpful when they appeal to their own demographic group. As seen in Figure 2a, personalized interventions that are also aligned are given a higher mean helpfulness score (= 2.90) than non-personalized ones (= 2.71) ($p < .05$).[5] Among personalized explanations, those that are sufficiently aligned with the user's identities are also perceived to be more helpful (= 2.90) than misaligned ones (= 2.61). However, since self-reported helpfulness may be unreliable, we also look at accuracy.

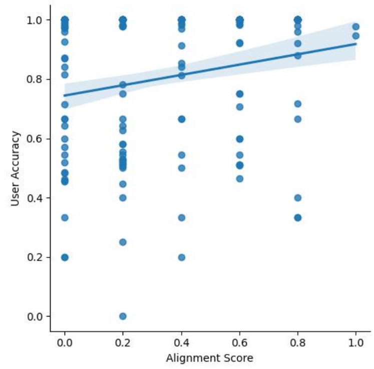
4.5 User Accuracy Results

To study the relationship between personalization and user accuracy, we recruit 157 participants with

[5]This is confirmed by both a standard t-test and Mann-Whitney U test.

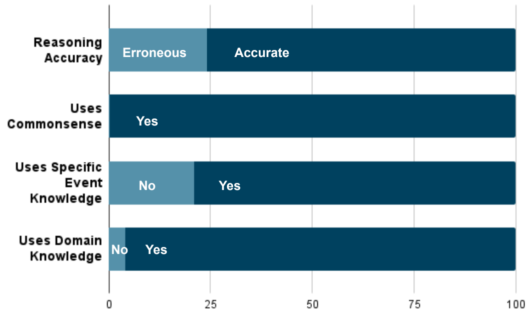


(a) Mean helpfulness scores for users receiving misaligned explanations (personalization alignment score 0.2 or lower, left), aligned explanations (alignment score 0.4 or higher, center left), explanations with alignment score of 0.6 (center right), and explanations without personalization (far right).

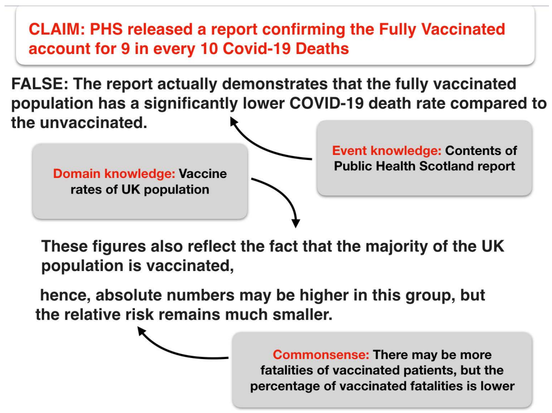


(b) Linear regression analysis with 95% confidence intervals showing explanation alignment to user attributes (x) and user accuracy (y) on a 0-1 scale.

Figure 2: Effects of personalization on self-reported helpfulness of explanations (left) and user accuracy (right).



(a) Breakdown of explanations by %.



(b) Example of reasoning types (commonsense, event knowledge, domain knowledge).

Figure 3: Analysis of all GPT-4 explanations. We provide a breakdown of explanation quality: (1) reasoning accuracy, (2) whether explanations use commonsense reasoning that should be innate to humans, (3) whether explanations require learned background knowledge from specific news events and (4) whether explanations require domain knowledge (e.g. scientific facts or knowledge of legal processes).

varying degrees of known alignment with personalized explanations for 150 claims. Figure 2b shows the linear regression analysis. For a threshold of 0.4, we find that user shown personalized explanations have an avg. accuracy of 85.89% vs. a non-personalized user accuracy of 76.65% (p=0.008).

5 How Well Do LLMs Explain?

We speculate that the strong performance described in the previous two sections is due in part to (1) the fact LLM-based interventions can convey more information than other scalable intervention types, and (2) the effectiveness of LLMs at retrieving relevant evidence from pretraining data without explicit instruction. In this section, we conduct an in-depth analysis of reasoning and form in LLM-generated explanations. We specifically address two safety risks that can arise from the way in which a language model argues for a given label. The first risk is that factually inaccurate or misleading content can arise in explanations from erroneous reasoning. The second risk is the potential for changes to linguistic properties of LLM explanations during personalization that may lead to stereotyping.

5.1 The Factuality Bottleneck

It should be noted that users' trust in AI is only beneficial if the model is accurate at label prediction. For the purposes of direct comparison across interventions in our study, we assume an oracle setting where the intervention label always matches the ground-truth label. We conduct a manual qualitative analysis of GPT-4 explanations generated

Group	Varied attribute	Avg. length (words) ↑	Avg. readability ↑	Avg. formality ↑
$g_{control}$	No personalization	52.59*	40.67*	92.63*
g_1	Default prompt: Conservative, White, Uneducated, Male, 30-49 years old	58.42	55.95	78.02
g_2	Liberal	58.45	55.99	77.84
g_3	Black	58.34	59.25*	71.42*
g_4	Educated	63.23*	38.37*	96.48*
g_5	Female	58.62	51.56*	87.81*
g_6	Age 65+	55.98*	55.04	81.67*

Table 3: Comparison of generic GPT-4 and personalized explanations across various demographic groups using automatic metrics. Higher scores indicate greater readability or formality respectively. Statistically significant differences between g_1 and g_k are marked by *.

for all true and false headlines (373 total headlines) with a Master's student who is not a co-author. We find that despite the oracle setting 24.13% of explanations use erroneous reasoning to validate claim labels. A detailed breakdown of reasoning used in explanations is provided in Figure 3. We see that the model is heavily reliant on memorized event knowledge (79.09% of explanations), which can be learned from web content like fact-checking articles and retrieved without explicit instruction from parametric knowledge. For actual deployment, steps will need to be taken to address issues of inaccurate prediction or model hallucination (Dziri et al., 2022; Guan et al., 2023; Kalai and Vempala, 2023) due to stored parametric knowledge being out-of-date. Some potential directions include use of retrieval-augmented generation approaches (Lewis et al., 2020; Nakano et al., 2021; Zhou et al., 2024).

5.2 Linguistic Effects of Personalization

We compare the average length, readability and formality of personalized explanations of the ground-truth label for six different demographic groups and GPT-4 generated explanations with no personalization, denoted $g_{control}$. The first group we consider, denoted as g_1, has the following demographic attribute values: *political affiliation* = ***conservative***, *race* = ***white***, *education* = ***uneducated***, *gender* = ***male***, *age* = ***30-49***. We then consider personalized explanations for additional groups that differ from g_1 by exactly one attribute. Specifically, for g_2 *political affiliation* = ***liberal***, for g_3 *race* = ***black***, for g_4 *education* = ***educated***, for g_5 *gender* = ***female***, and for g_6 *age* = ***65+***. For each of these demographics, we generate personalized explanations for the ground-truth label using prompts described in §3.1. We then measure differences between explanations across groups, using length, formality prediction (Pavlick and Tetreault, 2016), and reading difficulty based on the Flesch–Kincaid grade level metric (Flesch, 1948). Statistical significance is assessed using a standard 2-sided t-test.

From Table 3, we can see that lengths of explanations are relatively consistent across personalization settings. Political affiliation has the least effect across attributes, while readability and formality are significantly impacted by race, age, education and gender. In particular, specifying that the user is "educated" greatly reduces readability, indicating use of more challenging language, and increases formality by 18.46%. Specifying that the user is "black" leads to the least formal language usage. While not inherently harmful in this setting, it does indicate potentially discriminatory assumptions held by the model that are based on the user's demographics.

6 Conclusion

In conclusion, we introduced a framework for generating, personalizing and comprehensively evaluating LLM-based misinformation interventions (MisinfoEval). Our findings show a promising direction for social media platforms and policy makers to combat misinformation by improving the presentation of content to users. With the ability of LLMs to efficiently generate explanations to support veracity judgments with user personalization, they have the potential to serve as key components in designing scalable and powerful interventions. However, it is important to highlight that their success depends on model accuracy at predicting label veracity. This requires further improvements in automated prediction tools, greater coordination with human fact-checkers, or both. However, our observations raise concern that in the future advanced LLMs like GPT-4 can be misused to create targeted "fake news" campaigns against certain groups or even individuals. The personalization ability of LLMs is a double-edged sword, and collaboration between policy makers, researchers and engineers is needed to ensure they are used for ethical and desirable intentions.

7 Ethics Statement and Limitations

An important aspect of (mis)information diffusion is social network interactions. It has been shown that social cues and influence of connected users contribute to spread of false content (Avram et al., 2020; Starbird et al., 2023). The complexity of modeling these effects renders network interaction beyond the scope of our current study. However, we encourage future work that addresses whether explanations counteract detrimental network effects.

Beyond this limitation, another concern is misuse of LLM personalization by bad actors. While our work is focused on using LLM explanations to increase online literacy and mitigate effects of misinformation, they are a potential dual-use technology. There is a need for research focused on risks of personalization being exploited to generate more persuasive misinformation and manipulate public opinion, especially given the potential use of LLMs in political campaigning (Alvarez et al., 2023). As we show in §5, even in benign use cases, there are still risks of LLM deployment for content moderation like hallucinations and model demographic bias. We hope this work facilitates development of automated personalized content moderation that bears in mind risks of stereotyping or discrimination, which has occurred in other uses of personalized LLMs (Wan et al., 2023).

Acknowledgments

We thank David Rand for providing the data used in the study and Daniel Huttenlocher for thought-provoking discussions. We also thank colleagues at NYU, especially Julian Michael, Claudia Shi, Hannah Rose Kirk, Betty Hou, Jason Phang, and Salsabila Mahdi, for providing feedback on an early version of the study design, as well as the Dartmouth data analysis team (specifically Rong Guo) for help with summarizing experimental results. We are grateful to the anonymous ACL reviewers for their helpful comments and the MIT Generative AI Award committee for reviewing an abstract for the paper.

References

Daron Acemoglu, Asuman Ozdaglar, and James Siderius. 2023. A model of online misinformation. *Review of Economic Studies*.

R. Michael Alvarez, Frederick Eberhardt, and Mitchell Linegar. 2023. Generative ai and the future of elections.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. 2020. Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review*.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *ArXiv*, abs/2311.05656.

Katherine Clayton, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou, and Brendan Nyhan. 2020. Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, pages 1–23.

Valdemar Danry, Pat Pataranutaporn, Ziv Epstein, Matthew Groh, and Pattie Maes. 2022. Deceptive ai systems that give explanations are just as convincing as honest ai systems in human-machine decision making. *Presented at the International Conference on Computational Social Science (IC2S2)*, Extended Abstract.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Vittoria Elliott and David Gilbert. 2023. Elon musk's main tool for fighting disinformation on x is making the problem worse, insiders claim. *Wired*.

Ziv Epstein, Nicolò Foppiani, Sophie Hilgard, Sanjana Sharma, Elena L. Glassman, and David G. Rand. 2021. Do explanations increase the effectiveness of ai-crowd generated fake news warnings? In *International Conference on Web and Social Media*.

Rudolf Franz Flesch. 1948. A new readability yardstick. *The Journal of applied psychology*, 32 3:221–33.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127, Dublin, Ireland. Association for Computational Linguistics.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2023. Language models hallucinate, but may excel at fact verification. *ArXiv*, abs/2310.14564.

Brian M. Guay, Adam J. Berinsky, Gordon Pennycook, and David Rand. 2023. How to think about whether misinformation interventions work. *Nature Human Behaviour*, 7:1231 – 1233.

Yi-Li Hsu, Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. Is explanation the cure? misinformation mitigation in the short term and long term. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1313–1323, Singapore. Association for Computational Linguistics.

Farnaz Jahanbakhsh, Yannis Katsis, Dakuo Wang, Lucian Popa, and Michael Muller. 2023. Exploring the use of personalized ai for identifying misinformation on social media. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–27.

Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW2).

Chenyan Jia, Alexander Boltz, Angie Zhang, Anqing Chen, and Min Kyung Lee. 2022. Understanding effects of algorithmic vs. community label on perceived accuracy of hyper-partisan misinformation. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Skyler B Johnson, Henry S. Park, Cary P. Gross, and James B. Yu. 2017. Use of alternative medicine for cancer and its impact on survival. *Journal of the National Cancer Institute*, 110 1.

Adam Tauman Kalai and Santosh S. Vempala. 2023. Calibrated language models must hallucinate. *ArXiv*, abs/2311.14648.

Eleni Kyza, Christiana Varda, Loukas Konstantinou, Evangelos Karapanos, Serena Coppolino Perfumi, Mattias Svahn, and Yiannis Georgiou. 2021. Social media use, trust and technology acceptance: Investigating the effectiveness of a co-created browser plugin in mitigating the spread of misinformation on social media. *AoIR Selected Papers of Internet Research*.

Ro'ee Levy. 2021. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–70.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*.

Bingjie Liu. 2021. In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human–AI Interaction. *Journal of Computer-Mediated Communication*, 26(6):384–402.

Charles G. Lord, Lee D. Ross, and Mark R. Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37:2098–2109.

Lauren Lutzke, Caitlin Drummond, Paul Slovic, and Joseph Árvai. 2019. Priming critical thinking: Simple interventions limit the influence of fake news about climate change on facebook. *Global Environmental Change*, 58:101964.

Reid McIlroy-Young, Jon Kleinberg, Siddhartha Sen, Solon Barocas, and Ashton Anderson. 2022. Mimetic models: Ethical implications of ai that acts like you. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 479–490, New York, NY, USA. Association for Computing Machinery.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332.

Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175 – 220.

OpenAI. 2023. Gpt-4 technical report.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. 2019. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Political Communication eJournal*.

Gordon Pennycook, Jabin Binnendyk, Christie Newton, and David Rand. 2021a. A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, 7.

Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio Alonso Arechar, Dean Eckles, and David G. Rand. 2021b. Shifting attention to accuracy can reduce misinformation online. *Nature*, 592.

Francesco Pierri, Luca Luceri, Nikhil Jindal, and Emilio Ferrara. 2022. Propaganda and misinformation on facebook and twitter during the russian invasion of ukraine. *Proceedings of the 15th ACM Web Science Conference 2023*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Unpublished manuscript*.

Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri Jacques Geiss, Paolo Rosso, and Lucie Flek. 2022. FACTOID: A new dataset for identifying misinformation spreaders and political bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3231–3241, Marseille, France. European Language Resources Association.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *In Proceedings of ICLR*.

Haeseung Seo, Aiping Xiong, and Dongwon Lee. 2019. Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 265–274, New York, NY, USA. Association for Computing Machinery.

M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar. 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. *Commun. ACM*, 61(3):39–41.

Kate Starbird, Renée DiResta, and Matt DeButts. 2023. Influence and improvisation: Participatory disinformation during the 2020 us election. *Social Media + Society*, 9(2):20563051231177943.

Briony Swire-Thompson and David M. J. Lazer. 2019. Public health and online misinformation: Challenges and recommendations. *Annual review of public health*.

Ben M. Tappin, Gordon Pennycook, and David G. Rand. 2019. Bayesian or biased? analytic thinking and political belief updating. *Cognition*, 204.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9677–9705, Singapore. Association for Computational Linguistics.

Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Jingwen Zhang, Jieyu Featherstone, Christopher Calabrese, and Magdalena Wojcieszak. 2021. Effects of fact-checking social media vaccine misinformation on attitudes toward vaccines. *Preventive Medicine*, 145:106408.

Xinyi Zhou, Ashish Sharma, Amy X. Zhang, and Tim Althoff. 2024. Correcting misinformation on social media with a large language model.

A Appendix

A.1 Explanation Generation Details

For reproducibility, the specific GPT-4 model version we use is gpt-4-0613 with default API settings. The exact parameter count of GPT-4 is unknown.

A.2 Data Details

The news claim dataset used in this work was shared by the authors of (Pennycook et al., 2021a) and used as intended. The dataset is entirely in English. Since the original data only contained headline images, Python-tesseract was used to extract text headlines.[6]

A.3 Other Experimental Details

We used SciPy for statistical analysis in the paper (Virtanen et al., 2020).

A.4 Human Evaluation Setup

We received IRB exemption approval from the MIT Committee on the Use of Humans as Experimental Subjects (COUHES) for this study. We show screenshots of full instructions to crowdworkers in Figure 4. All participating crowdsource workers were compensated at a rate of between $0.40-$1 per example depending on the study length and qualification stage, which we determined to be a fair wage given best practices for compensation of crowdsource workers, the simplicity of task and estimated time commitment (Silberman et al., 2018). All participants explicitly consented to take part in the task after reading a short description. The user-claim interaction instance counts for Phase I are given in Table 4. These counts enable robust statistical analysis, in particular calculation of 95% confidence intervals shown in Table 2 with a 1-3.6% margin of error.

Intervention Type	Control	After
Label	1,349	1,171
Reaction Frame	2,181	1,540
GPT-4	742	604
Methodology AI	1,419	1,740
Methodology Human	2,272	1,655

Table 4: Instance counts pre-intervention (Control) and post-intervention (After) for all user-claim interactions in Table 2.

A.5 Personalization Details

Helpfulness Study. For the user-reported helpfulness personalization study, we select participants based on their inferred alignment with generated personalized explanations. We generate explanations before participant selection using the following attribute sets: [conservative, uneducated, male], [moderate, white, educated, female, 30-49], [moderate, white, educated, male, 30-49], [moderate, white, educated, male, 50-64], [moderate, white, uneducated, female, 30-49], [moderate, white, uneducated, female, 50-64].[7]

We infer user attributes by using the questionnaire in component (3) of the experiment to ask each user a list of Pew Research American Trends Panel[8] survey questions (Santurkar et al., 2023) on social and political issues in the United States. We then compute the conditional probability of a person with a set of demographic attribute values giving the same answers as the user. We choose the demographic group with the highest probability. The questionnaire in component (3) of the survey also asks for the actual demographic attributes of the user for validation.[9] Since they may not match the values of inferred attributes used to generate the explanation, we compute the personalization alignment score for each user, $s_{u_i e_j}$, defined as the proportion of attributes among those used to generate the personalized explanation e_j that are equal to the user u_i's self-reported ground-truth attributes.

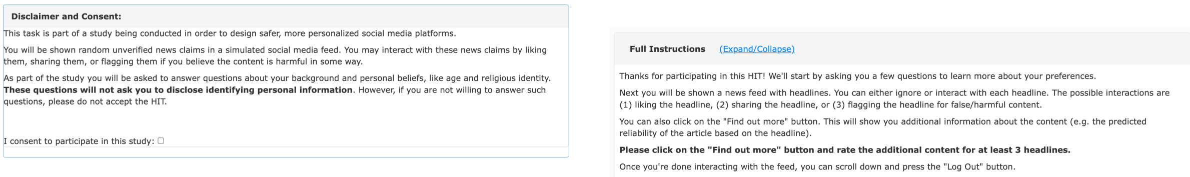
Accuracy Study. Participants are selected to ensure representation from left and right leaning self-reported ground-truth political ideologies, as well as unknown ideology participants. 54 participants are right-leaning and 23 are left-leaning. Explanations were personalized using the following at-

[6] https://pypi.org/project/pytesseract/

[7] Note that due to the less diverse spread of conservative workers in our experiment, predicted when inferring attributes with Pew Research survey data, we use 3 instead of 5 attributes during personalization (political affiliation, education and gender).

[8] https://www.pewresearch.org/our-methods/u-s-surveys/the-american-trends-panel/

[9] We decide to use the actual demographic values of users only for validation for several reasons. Real-world social media platforms often cannot obtain their exact values from the user or to use them in algorithms, especially for attributes like gender and race, due to privacy concerns or lack of information. This points to the need of inferring these values. Additionally, such a process allows us to perform a more fine-grained analysis on how personalization alignment affects the effectiveness of interventions.



Disclaimer and Consent:

This task is part of a study being conducted in order to design safer, more personalized social media platforms.

You will be shown random unverified news claims in a simulated social media feed. You may interact with these news claims by liking them, sharing them, or flagging them if you believe the content is harmful in some way.

As part of the study you will be asked to answer questions about your background and personal beliefs, like age and religious identity. **These questions will not ask you to disclose identifying personal information**. However, if you are not willing to answer such questions, please do not accept the HIT.

I consent to participate in this study: ☐

Full Instructions (Expand/Collapse)

Thanks for participating in this HIT! We'll start by asking you a few questions to learn more about your preferences.

Next you will be shown a news feed with headlines. You can either ignore or interact with each headline. The possible interactions are (1) liking the headline, (2) sharing the headline, or (3) flagging the headline for false/harmful content.

You can also click on the "Find out more" button. This will show you additional information about the content (e.g. the predicted reliability of the article based on the headline).

Please click on the "Find out more" button and rate the additional content for at least 3 headlines.

Once you're done interacting with the feed, you can scroll down and press the "Log Out" button.

Figure 4: Full Amazon Mechanical Turk Instructions.

tributes: [conservative, white, uneducated, male, 18-24] or [conservative, white, educated, male, 18-24]. 52 participants observed explanations for uneducated users and 105 observed explanations for educated users. We focus on right-learning personalization since prior research has found right-leaning users to be disproportionately targeted and involved in spread of misinformation (Sakketou et al., 2022; Jia et al., 2022; Pierri et al., 2022).

A.6 Demographics of Crowdworkers

We found that 2,018 workers recruited by the study answered a voluntary demographic questionnaire about age, gender, religion, politics, race and common sources for news. 52% of workers are 25-34 years old, 30% are 35-44 years old, 9% are 45-54 years old, 4% are 55-64 years old, 4% are 18-24 years old and 1% are over 65. 64% of participants are male and 36% are female.[10] 87% identify as Christian, 5% as Hindu, 3% as Jewish, 2% as Atheist, 2% as Muslim and 1% as spiritual or other. For political ideology, 29% of participants identify as right-leaning, 26% as moderate, 19% as left-leaning, 18% as very right-leaning, and 8% as very left-leaning. For race/ethnicity, 74% of participants identify as White, 20% as Asian, 2% as Black, 2% as Hispanic, and 2% as Native American or Pacific Islander. Most common news sources are Twitter (X), the New York Times, Breitbart, CNN, BBC News and Instagram.

[10]Less than 1% were non-binary or other.