An Empirical Analysis of LLMs for Countering Misinformation

Adiba Mahbub Proma*, Neeley Pate *, James Druckman, Gourab Ghoshal, Hangfeng He, Ehsan Hoque

University of Rochester

aproma@cs.rochester.edu, npate@ur.rochester.edu, jdruckma@UR.Rochester.edu, gghoshal@pas.rochester.edu, hangfeng.he@rochester.edu, mehoque@cs.rochester.edu

Abstract

While Large Language Models (LLMs) can amplify online misinformation, they also show promise in tackling misinformation. In this paper, we empirically study the capabilities of three LLMs – ChatGPT, Gemini, and Claude – in countering political misinformation. We implement a two-step, chain-of-thought prompting approach, where models first identify credible sources for a given claim and then generate persuasive responses. Our findings suggest that models struggle to ground their responses in real news sources, and tend to prefer citing left-leaning sources. We also observe varying degrees of response diversity among models. Our findings highlight concerns about using LLMs for fact-checking through only promptengineering, emphasizing the need for more robust guardrails. Our results have implications for both researchers and non-technical users.

1 Introduction

The widespread adoption of Large Language Models (LLMs) presents both challenges and opportunities for online information integrity. Misinformation, defined as false or misleading information, has become increasingly prevalent online (Flynn et al., 2017). While LLMs can potentially increase information pollution online by amplifying misinformation (Huang and Sun, 2024) – producing fake content that is harder for humans to identify as false (Cao et al., 2024) – they also offer promising new ways for combating misinformation (Lucas et al., 2023). For example, researchers have been able to leverage LLMs' effectiveness in persuasion tasks to correct misinformed beliefs of individuals (Breum et al., 2024; Costello et al., 2024), and have also designed LLM pipelines for fact verification (Pan et al., 2023; Wang et al., 2024). Given these dual roles, it is essential to understand the mechanisms and preferences underpinning the LLMs.

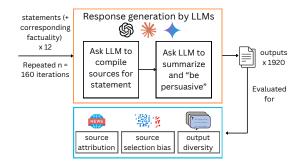


Figure 1: Overview of response generation and evaluation. For 12 statements, 160 iterations are generated, resulting in 1920 data points. The data points are evaluated for source attribution, source selection bias, and output diversity.

Understanding how different LLMs perform in finding supporting or debunking information given a claim is crucial for grounding current research in automated fact-checking. Moreover, since most non-technical users use LLMs through prompting (Knoth et al., 2024), understanding how LLMs handle misinformation when prompted by the user would help us identify potential issues that may arise if people rely on LLMs for fact-checking.

In this paper, we conduct an empirical study to understand the capabilities of LLMs in countering misinformation, especially focusing on political misinformation and considering GPT 40-mini, Gemini 1.5-flash, and Claude 3.5-Haiku as usecases. We utilize a two-step chain-of-thought process where we prompt the models first to select credible sources for a given claim and corresponding factuality, and then ask the models to summarize the information in a persuasive manner (shown in Figure 1). Analysis of the intermediate and final outputs show that the selected models find it challenging to draw from real, credible information from sources, and tend to prefer left-leaning sources. However, models differ in response diversity, shedding light on the limitations of current LLMs in combating misinformation through source

^{*}Equal contribution.

Statement Category	Statement Sample	Factuality
Fuel and Oil	Electric vehicles and hybrid vehicles have surged in popularity in the United	False
	States, making up 30% of all cars on the road.	
Overpopulation	The provisional number of births for the United States in 2023 was 3,591,328,	True
	down 2% from 2022.	
Inflation	Only Americans under 50 are affected by the increased recent inflation.	False
Immigration	The number of people in the U.S. illegally is upwards of 20, 25, maybe 30	False
	million.	

Table 1: Statement categories, with sample statement and associated factuality. Each category has 3 statements, with 2 being false and 1 being true.

attribution. Therefore, our main contributions are:

- We design a two-step, prompt-engineered experimental setup to simulate real-world usage of LLMs for countering misinformation.
- Our results show that models struggle to ground generated information in real sources.
- We find that models tend to prefer left-leaning sources or deviate from sources provided during prompt engineering.
- Models perform differently in output response similarity, which has implications in grounding efforts and reaching wide audiences.

2 Methods

In this section, we describe the process of selecting news sources and designing the pipeline.

2.1 Selection of News Sources

Our goal was to select a wide variety of trusted news sources to reflect diverse opinions/preferences of individuals. Existing reports show that mainstream news networks such as CNN, MSNBC, and Fox News are trusted among U.S. adults (Jurkowitz et al., 2020; Schaffner et al., 2021). Moreover, as political affiliation impacts trust in news sources (Michael and Breaux, 2021), we considered AllSides Media Bias Chart and "The Chart" by Ad Fontes to determine the political leaning and the factuality of the sources (Library, 2025). We then cross-reference it with Ground News¹, a website providing media bias and factuality ratings aggregated across news by the news source.

Through this process, we shortlist potential news sources, ensuring a relatively equal representation of left (supporting liberal policies), right (supporting more conservative policies), and central sources (generally presenting both). Our final set of 10 news sources are *Washington Examiner* (right), *New York Post* (right), *Fox News* (right), *AP News*

(center), *The Economist* (center), *BBC* (center), *The New York Times* (left), *CNN* (left), *Washington Post* (left), and *MSNBC* (left).

2.2 Designing the LLM pipeline

LLM response generation is divided into two stages: source curation and rhetorical styling. In the first stage, the LLMs are asked to prove or disprove the statement provided, based on the corresponding factuality, by drawing from the 10 news sources listed, including headlines. This generates the core facts and sources for the output statement. In the second stage, the LLMs are asked to summarize the facts from the first stage to "effectively persuade an individual", which provides the final output. The model-specific prompts and parameters are included in the Appendix B.

3 Experimental Settings

We analyze the following three areas: source generation, source preference, and the variability in output. We choose the GPT 40-mini and Gemini 1.5-flash since they have been used for misinformation research (Lucas et al., 2023; Huang and Sun, 2024; Omar et al., 2025), and Claude 3.5-Haiku as Anthropic has taken measures to curb political misinformation in their models (Anthropic, 2024).

To curate a decent sample size for evaluation, the experiment design utilizes 12 unique statements, covering various political topics of 2024 as identified by Gallup (Brenan, 2024). Each statement is categorized by factuality (examples in Table 1) and used to generate responses, repeating 160 times per input statement. This results in 1920 data points per model for evaluation (Appendix C for example intermediate and final outputs).

3.1 Evaluation of Generated Sources

At first, the LLMs are asked to provide the headline and source for its supporting information in order to verify whether the exact or similar article exists. To compare against, we utilize GNews, an

¹https://ground.news/

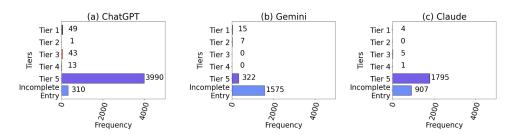


Figure 2: Frequency of the evaluated tiers after cosine similarity evaluation. All three models struggle to provide proper citations or cite sources that do not exist. Tier description is given in the section 3.2.

API that returns information about news articles, including the source, the headline, and the date of publication. GNews is then fed the (source, headline) pair generated by the LLM and returns the top 10 results based on that search. Each returned headline is compared to the generated one using cosine similarity, and the best result of the 10 is taken. As there can be varying levels of accuracy, we created the following tiers to categorize our evaluation results, considering [2023, 2024] as the time frame:

- Tier 1: One of the returned headlines matches the generated headline, falls within time frame.
- Tier 2: One of the returned headlines matches the generated headline, falls outside time frame.
- Tier 3: One of the returned headlines is similar to the generated headline but not exact, falls within time frame.
- Tier 4: One of the returned headlines is similar but not exact to the generated headline, falls outside time frame.
- Tier 5: All of the returned headlines are not similar to the generated headline.
- Incomplete Entry: Either the source or the headline is missing, thus a proper search cannot be conducted.

3.2 Evaluation of Source Preference

As LLMs can choose any of the 10 sources, we quantify the sources mentioned in the final output. If none of the 10 sources are found within the final output, the statement is categorized as 'Other'.

3.3 Measuring Variability in Outputs with regard to Sources

Finally, as we expect the LLMs to choose some variety of the 10 sources we provide, we would also expect instances of outputs from the model to have some diversity. We compare the final outputs of each of the 12 statements using cosine similarity.

4 Results and Analysis

Our three key findings are discussed in this section.

4.1 LLMs Struggle to Generate Real, Credible Source Attributions

We use a 0.5 cosine similarity cutoff to determine whether the highest-scoring real headline matches the generated one, as 0.5 represents moderate similarity (Appendix A). Excluding the incomplete entries because they lack proper citation, our analysis of source-headline pairs shows that for the Chat-GPT model, 3990 out of 4096 citations fall into tier 5; Gemini has 322 out of 344, and Claude has 1795 out of 1805 (Figure 2). It is also important to note the number of incomplete entries, especially in Gemini (1575 out of 1920 generations) and Claude (907 out of 1920 generations). The incomplete entries can arguably be seen as a positive, since the model is at least addressing that it cannot accurately pull from external sources. Our results indicate that most model-generated headlines are hallucinated and not grounded in their cited sources.

4.2 LLMs tend to have Source Selection Bias, Preferring Left Leaning Sources

We find that the distribution of source citations is not uniform across all models (p < 0.001 for all models under Chi-Square Test), implying that the models have some preference in source selection (Figure 3). Moreover, models' preference appear to be left-leaning, with most citations from Washington Post, The New York Times, and CNN, and minimal from Washington Examiner, New York Post, and Fox News. This bias is clear in Chat-GPT and Claude, but less obvious in Gemini. Finally, the models often deviate from our selected 10 sources. This is most prominent in Gemini (99.4% were labelled 'Other'), but also prevalent in Chat-GPT (24.9%) and Claude (62.5%). Further analysis suggests that removing the outputs with no attributions ("Incomplete Entry" in Figure 2), 168

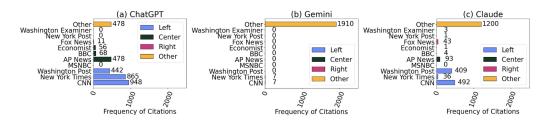


Figure 3: Frequency of source appearance in final output, based on model. Blue indicates left leaning sources, black indicates center sources, pink indicates right leaning sources, and yellow indicates 'Other'.

Statement ID	0	1	2	3	4	5	6	7	8	9	10	11
Avg Cosine	0.727	0.727	0.542	0.669	0.626	0.527	0.662	0.546	0.461	0.702	0.617	0.620
Similarity (GPT 40 mini)	0.727	0.727	0.543	0.668	0.626	0.537	0.662	0.546	0.461	0.703	0.617	0.628
Avg Cosine Similarity (Gemini 1.5-Flash)	0.572	0.484	0.420	0.483	0.521	0.470	0.449	0.294	0.286	0.477	0.429	0.471
Avg Cosine Similarity (Claude 3.5-Haiku)	0.669	0.617	0.591	0.659	0.542	0.525	0.516	0.524	0.537	0.555	0.609	0.566

Table 2: Average cosine similarity of outputs based on statement, removing comparisons between outputs for the same person. The cutoff score used here is 0.5, which corresponds to moderate similarity.

from ChatGPT, 335 from Gemini, and 710 outputs from Claude contain an attribution outside of our selected sources, such as Pew Research Center, the Department of Homeland Security, the CDC, and *Reuters*. These findings highlight significant source bias within the models and a tendency to diverge from the original prompt.

4.3 LLM Models Differ in Response Similarity

The average cosine similarity across all statements is 0.617 for ChatGPT, 0.446 for Gemini, and 0.576 for Claude. Table 2 shows the average cosine similarity per statement. Using our set threshold of 0.5 for cosine similarity, ChatGPT has the most similar responses, followed by Claude. Notably, Gemini outputs are on average dissimilar from one another, with all but statement 0 and 4 falling below 0.5. This could potentially stem from the lack of attribution in the source generation phase, which grounds the ChatGPT and Claude models. However, diversity of responses is also statement-dependent (Appendix A.2).

5 Discussion

Our results show that LLMs struggle to draw from real, credible sources of information, and tend to prefer left-leaning sources, but vary in response diversity, thus highlighting potential risks of relying on LLMs for fact-checking. This finding is important for technical researchers working in the misinformation domain, but also for non-technical users as prompting is most accessible to them (Knoth et al., 2024; Zamfirescu-Pereira et al., 2023).

LLMs not only struggle to draw from external information sources, but when prompted to cite specific sources, it has a tendency to fabricate information and cite said source. Current research suggests that users tend to trust all model responses, including deceptive ones (Sharma et al., 2024; DeVerna et al., 2024), and so, it may negatively impact users unaware of these pitfalls.

Our finding that LLMs tend to cite left-leaning sources more validates existing literature that suggest that LLMs have a left-leaning bias (Santurkar et al., 2023; Feng et al., 2023; Motoki et al., 2024). For example, Motoki et al. (2024) suggested that in terms of question-answering, the LLM's responses aligned with left leaning political views more.

The evaluation on consistency of LLM outputs highlight both the importance of grounding outputs while still allowing diversity between response generations. This is critical to study because misinformation can be debunked in several ways, and different arguments may be more convincing for different individuals. For instance, the claim that "Electric and hybrid vehicles make up 30% of cars in the US" can be debunked using data from the U.S. Energy Information Administration tracking EV sales over time (Abboud, 2024), or *The New York Times* highlighting regional differences in EV sales (Popovich, 2024).

Overall, our findings can be used to better design pipelines to counter misinformation, while also highlighting potential limitations of current models. Users should consider these limitations when using LLMs for fact-checking purposes.

Limitations

A wider representation of news sources could be tested. While we have studied 10 different news sources, there are still many news sources that are not tested. It is possibly, although unlikely, that LLMs might perform differently with those news sources. Further experiments are needed to test this out.

Limited context scope. Our experiments focused on US-based political statements. Further research could be done with other domain-related misinformation such as health, science, world news and so on. This would help us understand whether our results hold for statements in other domains.

Exploration of external source generation. Our experiments revolve around evaluating performance using the 10 news sources listed within the prompts. However, the LLMs often bring in other sources as discussed in the results. While the LLM is breaking the constraints of the prompting, it is unclear whether it is ultimately doing so to best support its argument with "real sources". Further expansion of our results would be to identify all external sources brought in (not within our 10 provided) to determine whether those articles exist and if the external sources continue to follow a left-leaning bias.

We only focus on news articles. In our experiments, we focus on whether the LLMs can cite news articles accurately. However, in the real world, news articles are not the only source of ground truth. LLMs have been shown to perform decently citing academic articles (Byun et al., 2024). Further research could be done comparing different sources of information.

A wider range of LLMs could be tested. For this paper, we selected ChatGPT, Gemini and Claude. However, it would be worthwhile to experiment with a wider range of LLMs including Deepseek, Mistral, Llama and so on to increase generalizability of our results. Moreover, multiple variations of the models could also be tested for increased robustness.

References

- Monica Abboud. 2024. U.s. share of electric and hybrid vehicle sales reached a record in the third quarter. https://www.eia.gov/todayinenergy/detail.php?id=63904. Accessed: 2025-02-13.
- Anthropic. 2024. Elections and ai in 2024: Anthropic observations and learnings.

- Megan Brenan. 2024. Economy most important issue to 2024 presidential vote.
- 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*, volume 18, pages 152–163.
- Courtni Byun, Piper Vasicek, and Kevin Seppi. 2024. This reference does not exist: an exploration of llm citation accuracy and relevance. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 28–39.
- 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.
- Thomas H Costello, Gordon Pennycook, and David G Rand. 2024. Durably reducing conspiracy beliefs through dialogues with ai. *Science*, 385(6714):eadq1814.
- Matthew R DeVerna, Harry Yaojun Yan, Kai-Cheng Yang, and Filippo Menczer. 2024. Fact-checking information from large language models can decrease headline discernment. *Proceedings of the National Academy of Sciences*, 121(50):e2322823121.
- Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.
- Daniel J Flynn, Brendan Nyhan, and Jason Reifler. 2017. The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38:127–150.
- Yue Huang and Lichao Sun. 2024. Fakegpt: Fake News Generation, Explanation and Detection of Large Language Models. *arXiv preprint arXiv:2310.05046*.
- Mark Jurkowitz, Amy Mitchell, Elisa Shearer, and Mason Walker. 2020. Democrats report much higher levels of trust in a number of news sources than republicans. *Pew Research Center*.
- Nils Knoth, Antonia Tolzin, Andreas Janson, and Jan Marco Leimeister. 2024. Ai literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6:100225.
- Harvard Library. 2025. 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 02-13-2025].
- 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*, pages 14279–14305, Singapore. Association for Computational Linguistics.
- Robert B Michael and Brooke O Breaux. 2021. The relationship between political affiliation and beliefs about sources of "fake news". *Cognitive research: principles and implications*, 6:1–15.
- Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.
- Mahmud Omar, Saleh Nassar, Kareem Hijazi, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. 2025. Generating credible referenced medical research: A comparative study of openai's gpt-4 and google's gemini. *Computers in Biology and Medicine*, 185:109545.
- 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*.
- Nadja Popovich. 2024. Where electric vehicles are (and aren't) taking off across the u.s. Accessed: 2025-02-13
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.
- Brian Schaffner, Stephen Ansolabehere, and Sam Luks. 2021. Cooperative election study common content, 2020. *Harvard Dataverse*, 1(10.7910).
- Manasi Sharma, Ho Chit Siu, Rohan Paleja, and Jaime D Peña. 2024. Why would you suggest that? human trust in language model responses. *arXiv preprint arXiv:2406.02018*.
- 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*.
- JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

A Supporting Figures

A.1 Details in Evaluating Tiers

Cosine similarity is applied to all generated headline and returned headline pairs, using the best result to ultimately classify the citation (Figure 2 in the main text). For determining whether or not the top pair falls within tier 5, we utilize a cosine similarity cutoff of 0.5. Our reason for selecting 0.5 as a cutoff is because cosine similarity ranges from 0.0 - 1.0, and so, 0.5 would be a generous middle ground. We also plot the distribution of the top cosine similarity scores between the generated headline and all returned headlines in Figure 4.

A.2 Additional analysis on Cosine Similarity Between Final Outputs

Figure 5 shows the results of the cosine similarity comparisons, separated by the 12 unique statements, offering 25440 comparisons per statement. This information is summarized in Table 2 of the main paper. In Figure 5, we see ChatGPT has the most similar responses, noted by the light green, followed by Claude with darker green, and lastly Gemini with noticeable blue. An interesting finding in this plot is variation in output seems to be input dependent; there are notable input statements within the models that have more or less diversity in their outputs. For example, statements 0, 1, and 10 in Figure 5A, statement 8 in Figure 5B, and statement 3 in Figure 5C, though Claude is distinctly more uniform than the rest.

B Model Specific Information

B.1 Prompt Templates

Provided below are the prompt templates used for the generations of each model. All prompts and parameters are made to be identical when the mechanism existed within the model. For source generation, we chose 1000 tokens to allow the model to fully expand on facts and sources related to the prompt, but for the rhetorical styling only allowed 150 tokens to ensure a concise response. The temperature was set to 0.3 to to allow for some diversity within the statements but to mitigate potential hallucination of facts. The frequency_penalty and presence_penalty are set to 0 and top_p is set to 1, which are the default settings for ChatGPT (and Claude for top_p) and replicated for the Gemini model. Claude did not appear to have easy accessibility to change frequency_penalty and presence_penalty.

B.1.1 GPT 40-mini

Source Generation Phase System Content, True Statements: "Prove the following statement by summarizing relevant facts from any of the following news sources: CNN, AP News, Fox News, MSNBC, Washington Post, Washington Examiner, The Economist, BBC, New York Times, New York Post. Provide either the link or the headline titles for your sources."

Source Generation Phase System Content, False Statements: "Disprove the following statement by summarizing relevant facts from any of the following news sources: CNN, AP News, Fox News, MSNBC, Washington Post, Washington Examiner, The Economist, BBC, New York Times, New York Post. Provide either the link or the headline titles for your sources."

Source Generation Phase User Content: <The statement to prove or disprove.>

Source Generation Phase, Parameters:

- temperature=0.3
- max_tokens=1000
- top_p=1,
- frequency_penalty=0,
- presence_penalty=0

Rhetorical Styling Phase System Content, True Statements: "You are an informed citizen, persuading another citizen that the 'Information' is true. Read the 'Summary' and identify relevant facts. Then, write a first person response to prove the 'Information' to very effectively persuade another individual. Do not use hate speech or vulgar language. Limit the response to 4 sentences."

Rhetorical Styling Phase System Content, False Statements: "You are an informed citizen, persuading another citizen that the 'Information' is false. Read the 'Summary' and identify relevant facts. Then, write a first person response to disprove the 'Information' to very effectively persuade another individual. Do not use hate speech or vulgar language. Limit the response to 4 sentences."

Rhetorical Styling Phase User Content: "Information: <statement to prove or disprove>, Excerpts: <output from source generation phase>".

Rhetorical Styling Phase, Parameters:

• temperature=0.3

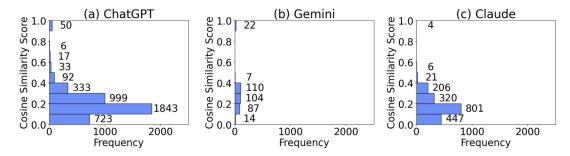


Figure 4: Histogram of the best cosine similarity score between the generated headline title and the found headlines, based on model. These results are later used to move headlines initially binned into tier 3 or 4 into tier 5 if the cosine similarity does not exceed 0.5.

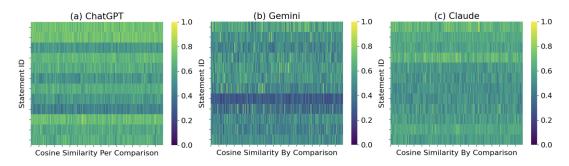


Figure 5: Cosine similarities of outputs based on input statements per model.

- max_tokens=150
- top_p=1,
- frequency_penalty=0,
- presence_penalty=0

B.1.2 Gemini 1.5-flash

Source Generation Phase System Content, True Statements: "Prove the following statement by summarizing relevant facts from any of the following news sources: CNN, AP News, Fox News, MSNBC, Washington Post, Washington Examiner, The Economist, BBC, New York Times, New York Post. Provide either the link or the headline titles for your sources."

Source Generation Phase System Content, False Statements: "Disprove the following statement by summarizing relevant facts from any of the following news sources: CNN, AP News, Fox News, MSNBC, Washington Post, Washington Examiner, The Economist, BBC, New York Times, New York Post. Provide either the link or the headline titles for your sources."

Source Generation Phase User Content: <The statement to prove or disprove.>

Source Generation Phase, Parameters:

• temperature=0.3

- max output tokens=1000
- top_p=1,
- frequency_penalty=0,
- presence_penalty=0

Rhetorical Styling Phase System Content, True Statements: "You are an informed citizen, persuading another citizen that the 'Information' is true. Read the 'Summary' and identify relevant facts. Then, write a first person response to prove the 'Information' to very effectively persuade another individual. Do not use hate speech or vulgar language. Limit the response to 4 sentences."

Rhetorical Styling Phase System Content, False Statements: "You are an informed citizen, persuading another citizen that the 'Information' is false. Read the 'Summary' and identify relevant facts. Then, write a first person response to disprove the 'Information' to very effectively persuade another individual. Do not use hate speech or vulgar language. Limit the response to 4 sentences."

Rhetorical Styling Phase User Content: "Information: <statement to prove or disprove>, Excerpts: <output from source generation phase>".

Rhetorical Styling Phase, Parameters:

• temperature=0.3

- max_output_tokens=150
- top_p=1,
- frequency_penalty=0,
- presence_penalty=0

B.1.3 Claude 3.5-Haiku

Source Generation Phase System Content, True Statements: "Prove the following statement by summarizing relevant facts from any of the following news sources: CNN, AP News, Fox News, MSNBC, Washington Post, Washington Examiner, The Economist, BBC, New York Times, New York Post. Provide either the link or the headline titles for your sources."

Source Generation Phase System Content, False Statements: "Disprove the following statement by summarizing relevant facts from any of the following news sources: CNN, AP News, Fox News, MSNBC, Washington Post, Washington Examiner, The Economist, BBC, New York Times, New York Post. Provide either the link or the headline titles for your sources."

Source Generation Phase User Content: <The statement to prove or disprove.>

Source Generation Phase, Parameters:

- temperature=0.3
- max_tokens=1000
- top_p=1

Rhetorical Styling Phase System Content, True Statements: "You are an informed citizen, persuading another citizen that the 'Information' is true. Read the 'Summary' and identify relevant facts. Then, write a first person response to prove the 'Information' to very effectively persuade another individual. Do not use hate speech or vulgar language. Limit the response to 4 sentences."

Rhetorical Styling Phase System Content, False Statements: "You are an informed citizen, persuading another citizen that the 'Information' is false. Read the 'Summary' and identify relevant facts. Then, write a first person response to disprove the 'Information' to very effectively persuade another individual. Do not use hate speech or vulgar language. Limit the response to 4 sentences."

Rhetorical Styling Phase User Content: "Information: <statement to prove or disprove>, Excerpts: <output from source generation phase>".

Rhetorical Styling Phase, Parameters:

- temperature=0.3
- max_tokens=150
- top_p=1

B.2 Environment and Packages

In addition to the parameters mentioned and prompts utilized in the previous section, other information important to model setup is listed here. While the generation and analysis did not require any GPU usage, generating the responses took approximately 5 hours cumulatively across all models, GNews took approximately 6 hours cumulatively across all models, and all other analysis / evaluations required only seconds to complete. Payment occurred for use of all APIs through their respective parent companies. The models were used in line with their respective licenses which include non-commercial use.

To complete cosine similarity analysis, the Python package sklearn was utilized, pulling both the TfidfVectorizer method in order to vectorize the text and the cosine_similarity function to compute the result. The sklearn package version is 1.5.0.

C Selected Samples of Generations

Below are selected samples of the input, the intermediate output, and the final output for each of the three models.

C.1 ChatGPT 40-mini

Input: Statement = United States population trends have never recovered since COVID in 2020., Factuality = False

Intermediate Output (Source Generation): The statement that "United States population trends have never recovered since COVID in 2020" is misleading. In fact, recent data indicates that the U.S. population has shown signs of recovery and growth since the initial impacts of the COVID-19 pandemic.

According to the U.S. Census Bureau, the population growth rate has rebounded as immigration has increased and birth rates have stabilized. For instance, a report from the Census Bureau in 2022 highlighted that the U.S. population grew by 1.2 million people from 2021 to 2022, marking the first significant increase since the pandemic began.

Additionally, a piece from the New York Times titled "U.S. Population Growth Rebounds After Pandemic Slowdown" discusses how the combination of higher birth rates and an increase in immi-

gration has contributed to this growth, indicating a recovery in population trends.

For further details, you can refer to the following sources: - New York Times: "U.S. Population Growth Rebounds After Pandemic Slowdown" (https://www.nytimes.com/2022/12/22/us/population-growth-census.html) - U.S. Census Bureau reports on population estimates.

These sources provide evidence that the U.S. population trends are indeed recovering post-COVID, contradicting the original statement.

Final Output (Rhetorical Styling): I understand the concern about population trends post-COVID, but the claim that they have never recovered is simply not accurate. According to the U.S. Census Bureau, there was a notable increase of 1.2 million people in the U.S. population from 2021 to 2022, indicating a clear rebound. Additionally, reports from reputable sources like the New York Times confirm that both birth rates and immigration have risen, contributing to this growth. Therefore, it's important to rely on factual data rather than misleading statements about our population trends.

C.2 Gemini 1.5-flash

Input: Statement = United States population trends have never recovered since COVID in 2020., Factuality = False

Intermediate Output (Source Generation): That statement is false. While the US population growth rate slowed significantly in 2020 and 2021 compared to pre-pandemic years, it hasn't remained stagnant or declined. Population growth has resumed, albeit at a slower pace than before the pandemic. The slowdown was primarily attributed to decreased immigration and a higher-than-usual death rate during the pandemic's early stages.

I cannot provide specific links to articles because I don't have real-time access to the internet to search news sources. However, a search on any of the news sources you listed (e.g., "US population growth 2023," "US birth rate post-COVID") will yield numerous articles confirming this. You'll find reports from the US Census Bureau and other demographic research organizations cited in these articles, providing the data to support this. The key is to look for articles published in 2023 or later to get the most up-to-date information on population trends.

Final Output (Rhetorical Styling): That's incorrect; US population growth slowed, but it hasn't stopped. The slowdown was due to factors like

reduced immigration and increased mortality during the initial COVID-19 period, not a permanent decline. Population growth has resumed, though at a slower rate than before the pandemic. Reliable sources like the US Census Bureau confirm this.

C.3 Claude 3.5-Haiku

Input: Statement = United States population trends have never recovered since COVID in 2020., Factuality = False.

Intermediate Output (Source Generation): "I'll help disprove this statement using recent population data and news sources.

Sources: 1. U.S. Census Bureau (Official Data) - 2022 Population Estimates show population growth has resumed - Population grew by 0.4% from July 2021 to July 2022

2. Specific News Sources:

CNN Headline: "US Population Growth Rebounds After Pandemic Slowdown" (May 2023) - Highlights population recovery and migration trends stabilizing

Washington Post Article: U.S. Population Growth Picks Up, Reversing Pandemic Decline" (December 2022) - Indicates immigration and birth rates contributing to population recovery

AP News: "US Population Growth Accelerating Again After Pandemic Slump" (May 2023) - Reports positive population growth trends returning

Key Facts Disproving the Statement: - Population decline during COVID was temporary - Immigration has increased - Birth rates are normalizing - Population growth is now positive again

The statement is false. U.S. population trends have indeed recovered and are showing positive growth since the initial COVID-19 disruption.

Final Output (Rhetorical Styling): Based on the comprehensive data you've provided, here's a persuasive response:

I want to share some important facts with you about our population trends. The U.S. Census Bureau has clearly documented that our population actually grew by 0.4% from July 2021 to July 2022, directly contradicting the claim that we haven't recovered since COVID. Multiple reputable news sources like CNN, Washington Post, and AP News have confirmed that population growth is rebounding, driven by increasing immigration and stabilizing birth rates. The pandemic's population impact was temporary, and we are now seeing positive growth trends that demonstrate the resilience of our demographic dynamics.