

Truth Sleuth & Trend Bender

AI Agents to fact-check YouTube videos & influence opinions

Cécile Logé

Department of Computer Science
Stanford University
ceciloge@stanford.edu

Rehan Ghori

SCPD NDO Student
Stanford University
rghori@stanford.edu

Abstract

Misinformation poses a significant threat in today’s digital world, often spreading rapidly through platforms like YouTube. This paper introduces a novel approach to combating misinformation by developing an AI-powered system that not only fact-checks claims made in YouTube videos but also actively engages users in the comment section and challenge misleading narratives. Our system comprises two main agents: Truth Sleuth and Trend Bender.

Truth Sleuth extracts claims from a YouTube video, uses a Retrieval-Augmented Generation (RAG) approach - drawing on sources like Wikipedia, Google Search, Google FactCheck - to accurately assess their veracity and generates a nuanced and comprehensive report. Through rigorous prompt engineering, Trend Bender leverages this report along with a curated corpus of relevant articles to generate insightful and persuasive comments designed to stimulate a productive debate. With a carefully set up self-evaluation loop, this agent is able to iteratively improve its style and refine its output.

We demonstrate the system’s capabilities through experiments on established benchmark datasets and a real-world deployment on YouTube, showcasing its potential to engage users and potentially influence perspectives. Our findings highlight the high accuracy of our fact-checking agent, and confirm the potential of AI-driven interventions in combating misinformation and fostering a more informed online space.

1 Introduction

Misinformation is one of the most pressing threats of our time, and YouTube videos serve as a major platform through which it can spread (IFCN, 2022). On top of this, the comment sections of these videos can become echo-chambers that amplify or reinforce misleading or harmful claims. Providing fact-checked information to address mis-

leading content has been shown to be more effective than simply removing it (Ecker et al., 2020). Diversifying the viewpoints users are exposed to can burst the filter bubble and get them out of their intellectual isolation.

Based on these two elements, our goal is to build an application that takes a YouTube video as input and not only fact-checks the claims made in the video but also outputs a comment to interact with users about these claims, protect them from the dangers of misinformation and ultimately maybe even change their minds. We chose two main themes for our experiments: the Manosphere, and Diet Culture. See Figure 1 for more insights about these two spaces.

2 Related Work

Fact-checking: The work by Polat et al., 2024 details several prompt engineering strategies e.g. incorporating a simple instruction accompanied by a task demonstration, and shows how crafting the right prompt can significantly enhance performance when it comes to using LLMs for knowledge extraction. It will provide great inspiration when extracting claims from the video transcripts in our project. On top of this, the work on Data Commons by Guha et al., 2023 and Google’s DataGemma models for fact-checking by Radhakrishnan et al., 2024 have shown a lot of promises and can guide the integration of external data sources in validating a YouTube video’s claims, helping to classify them as true, false, or unsure. These approaches exemplify the growing importance of Retrieval-Augmented Generation (RAG) techniques in grounding LLM outputs and enhancing their factual accuracy (Lewis et al., 2020), particularly in tasks like fact-checking. Finally, Akhtar et al., 2023 provided a broader overview of multi-modal fact-checking techniques that integrate both text and visuals to verify information, refining the

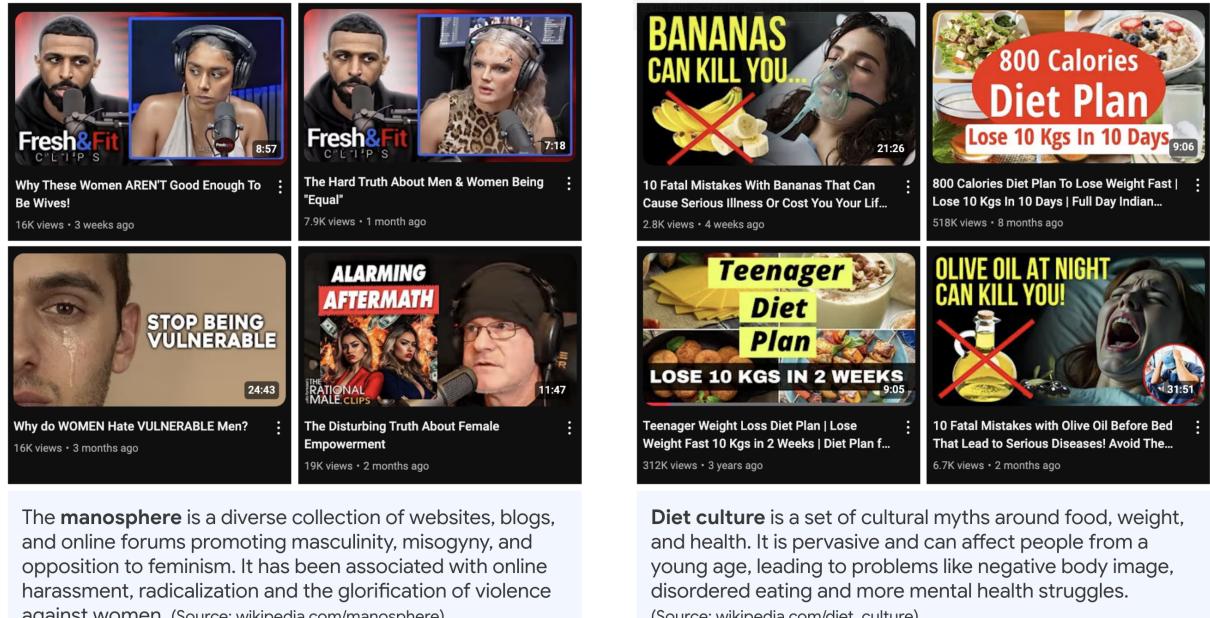


Figure 1: We chose two main themes for our experiments: the Manosphere and Diet Culture. Video screenshots from YouTube channels: Fresh&Fit Clips, The Rational Male, 10 SECRETS, Eat More Lose More.

fact-checking approach to take into account visual cues.

Perspectives & Persuasion: The idea of providing different viewpoints on a given topic while maintaining factual integrity has obvious parallels with the work by Shao et al., 2024 on STORM. Hayati et al., 2023 also experimented with generating diverse perspectives on subjective topics using LLMs and introduced an interesting criteria-based prompting technique to ground diverse opinions. The goal of convincing a reluctant or skeptical audience and/or changing their minds calls up the work by Furumai et al., 2024 on persuasive chatbots (PersuaBot). Their approach - from combining different persuasive strategies to selecting metrics to assess the quality of the conversation - is an invaluable source of inspiration.

Our project builds on cutting-edge research on large language models, prompt engineering, fact-checking, persuasion and multi-perspective generation, with the goal of combining them all to offer a unique approach to addressing misinformation on YouTube, helping users break free from echo chambers and bringing down ideological barriers.

3 Core Concepts & Architecture

Driven by the understanding that misinformation thrives in echo chambers and that countering it requires both factual correction and persuasion - two highly different tasks that relies on different sets of skills and logic - our system employs two primary agents working hand in hand:

- Our **Truth Sleuth** agent focuses on identifying and debunking misleading claims within a given YouTube video.
- Our **Trend Bender** agent generates a general comment (or a reply to a specific user) to be posted in the comment section of the video.

See Figure 2 for a general overview of the system.

3.1 Truth Sleuth: our Fact-Checking agent

Our Truth Sleuth agent is based on solid prompt engineering and API calls to reliable sources. To assess the veracity of claims, it employs a Retrieval-Augmented Generation (RAG) approach, and operates through the following pipeline:

1. **Information Extraction:** Through the YouTube API, relevant information from the video, including the title, channel information, and raw audio captions, is extracted. The agent starts by reformatting the audio captions and editing them appropriately for punctuation and grammar.

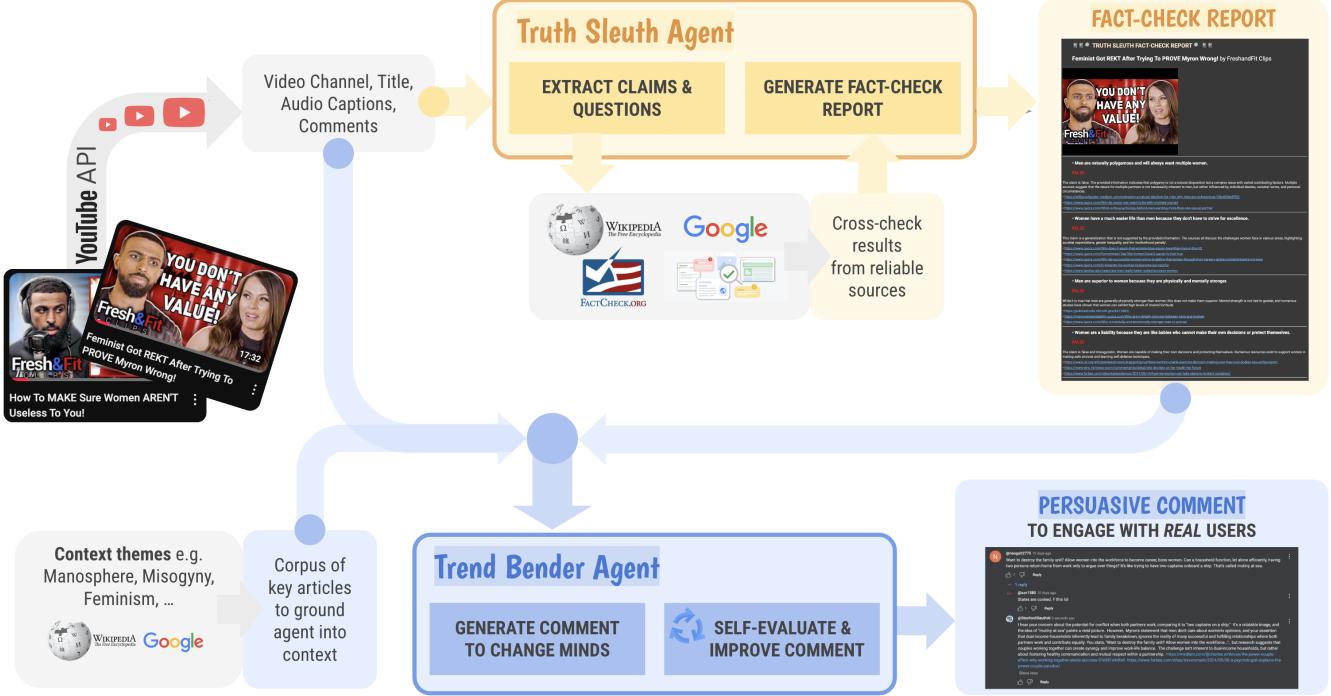


Figure 2: General overview of our system - comprising two main agents: **Truth Sleuth** and **Trend Bender**. **Truth Sleuth** extracts claims from a YouTube video, relies on APIs like Google Search, Wikipedia, Google FactCheck to accurately assess their veracity and generates a nuanced and comprehensive report. **Trend Bender** leverages this report and a curated corpus of relevant articles, along with a self-evaluation loop to generate insightful and persuasive comments designed to stimulate a productive debate.

2. Claim Extraction & Question Generation:

The agent then analyzes the extracted information to pinpoint specific claims made in the video. These claims are then transformed into precise questions suitable for fact-checking. *For example, if the video claims "Men are superior to women because they are physically and mentally stronger," the agent might generate questions like "Are men physically stronger than women?" and "Are men mentally stronger than women?"* The output is a JSON object that can be easily exploited over next steps.

3. Claim Assessment:

Questions are sent to the Google Search API ([Google for Developers, 2024](#)), the Wikipedia API ([Wikimedia Foundation, 2024](#)) and the Google FactCheck ClaimReview API ([Google News Initiative, 2024](#)). The agent analyzes the results, gathers evidence in the form of urls and excerpts, and cross-references them to assess the veracity of the identified claims. This process culminates in the generation of a JSON object containing the claim, verdict, reasoning, and supporting

source URLs. The verdict can be either *True*, *Partly True*, *Partly False*, *False* or *Unsure*.

4. Fact-Check Report:

Finally, a comprehensive and nuanced fact-check report is generated. It is rendered in both raw text format to be sent to the Trend Bender agent, and human-readable Markdown format (including the video thumbnail, colored indicators for verdicts) to be shared directly with the user.

In their experiments on headline discernment, [DeVerna et al., 2024](#) showed that LLM fact checks can actually increase belief in dubious headlines when the AI is unsure about a claim's veracity. For this reason, and to avoid any confusion and misinterpretation, claims deemed *Unsure* are automatically excluded from the final report.

See Figure 3 for a visual diagram of the Truth Sleuth agent.

3.2 Trend Bender: our Influencing agent

We have run several experiments on our Trend Bender pipeline, and concluded our best agent should rely on insights from the Truth Sleuth agent, context from a carefully-curated corpus of articles re-

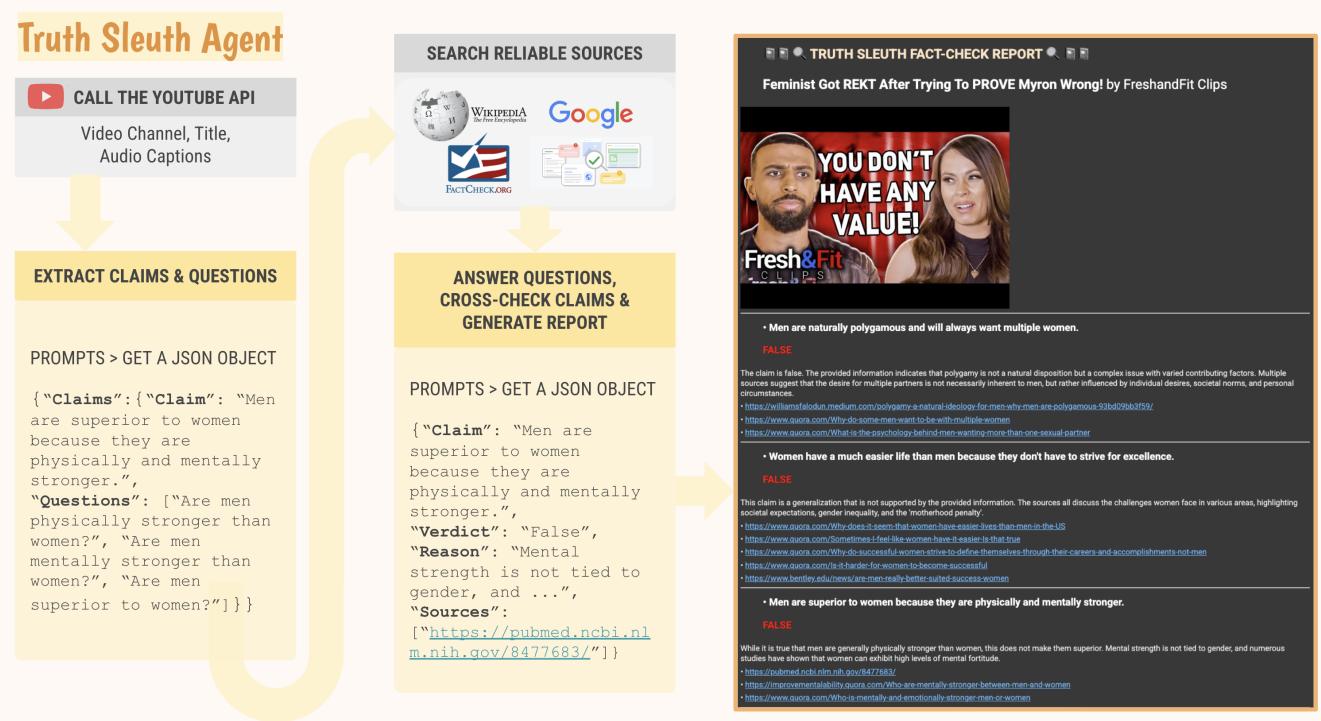


Figure 3: **Truth Sleuth** extracts claims from a YouTube video, uses a Retrieval-Augmented Generation (RAG) approach - drawing on Wikipedia, Google Search, Google FactCheck - to accurately assess their veracity and generates a nuanced and comprehensive report. Actual extract of a Fact-Check report shown on the right, for a video from the FreshandFit Clips channel.

lated to the video’s theme (e.g., Manosphere, Diet Culture) and a self-evaluation loop allowing it to fine-tune its output on its own:

- Information Extraction:** Similar to the Truth Sleuth pipeline, relevant information from the video, including the title, channel information, and audio captions, is gathered through the YouTube API. User comments are also extracted to provide information on tone and opinions. On top of this, the agent also leverages the Truth Sleuth fact-check report, and draws upon a curated corpus providing deeper contextual understanding.
- First Comment Generation:** Based on very detailed instructions, the agent generates a relevant and informed comment designed to engage users. If a specific comment has been highlighted as the one to reply to, it focuses on answering that user’s points and engaging them in a respectful discussion.
- Self-Evaluation & Improvement:** The agent evaluates its own comment based on a provided set of seven criteria, with scores from 0 to 2, and feedback on each rubrics. Using

this evaluation, it generates an improved comment now ready to be posted in the comment section of the YouTube video.

See Figure 4 for a visual diagram of the Trend Bender agent. See Section 4 for more details on the experiments, detailed prompts, and instructions.

4 Experiments & Results

Note that throughout all our experiments, we use Gemini as our Base LLM with the model `gemini-1.5-flash`.

4.1 Truth Sleuth Performance

To assess the performance of the Truth Sleuth agent, we conducted an experiment using established benchmark datasets: FEVER (Thorne et al., 2018) and AVeriTec (Schlichtkrull et al., 2024). Note that FEVER focuses on claims derived from Wikipedia, while AVeriTec comprises real-world claims from diverse sources, presenting a much more challenging evaluation scenario.

We used a subset of 105 claims - 50 from FEVER, 55 from AVeriTec, excluding "Not enough info" or "Conflicting evidence" claims -

Figure 4: Through rigorous prompt engineering, **Trend Bender** leverages this report along with a curated corpus of relevant articles to generate insightful and persuasive comments designed to stimulate a productive debate. With a carefully set up self-evaluation loop, this agent is able to iteratively improve its style and refine its output.

Datasets	FeVER	AVeriTec	Overall	w/o "Unsure"
ACCURACY	90%	73%	81%	93%
PRECISION				
on True/Supports	95%	78%	90%	90%
on False/Refutes	96%	94%	95%	95%
RECALL				
on True/Supports	91%	54%	77%	90%
on False/Refutes	89%	79%	83%	95%
F1 SCORE				
on True/Supports	93%	64%	83%	90%
on False/Refutes	93%	86%	89%	95%

Figure 5: Performance of the **Truth Sleuth** agent on a subset of 105 claims - 50 from FEVER, 55 from AVeriTec. The column "w/o Unsure" excludes cases where the agent's verdict is *Unsure* - never shown in the final report.

to evaluate the agent's ability to accurately classify claims as "Supports" (*True*, *Partly True*) or "Refutes" (*False*, *Partly False*). We chose classic metrics such as accuracy, precision, recall and F1 score. Results can be found in Figure 5.

Our agent achieved an overall accuracy of 81% on the combined dataset. When excluding cases

where the agent was *Unsure* (which are never shown in the final report), the accuracy increased to 93%. As expected, performance was lower on the more challenging dataset (AVeriTec), highlighting the complexity of real-world fact-checking.

For further reference, in their experiments, Schlichtkrull et al., 2024 mentions a best F1 score on the full AVeriTec dataset of 62% on "Supports" - surpassed by Truth Sleuth at 64% - and of 74% on "Refutes" - surpassed again at 86%.

4.2 Trend Bender Prompt Engineering

Having established the general architecture of the Trend Bender Agent, we conducted experiments to evaluate its effectiveness in generating persuasive and insightful comments. Specifically, we wanted to assess how different prompting strategies and input combinations influence the quality of the generated comments.

To conduct our human evaluation, we used a set of seven criteria - each scored on a scale of 0 to 2, with 2 being the highest - as follows:

- **Does Not Hallucinate:** Trend Bender - being an AI agent - should not pretend to be human or to have had human experiences. Additionally, it should not make up facts or sources.

and should rely strictly on the provided data when making factual claims.

- **Takes the 'Right' Stand:** The generated comment should not blindly agree, congratulate or praise content that is false or harmful.
- **Is Specific:** The generated comment should address specific points from the video, and can even quote some passages.
- **Is Sound & Logical:** The generated comment should not include logical fallacies or faulty phrasings.
- **Cites Evidence:** The generated comment should leverage the urls / sources from the fact-check report and themed corpus to back its position.
- **Avoids Truisms:** The generated comment should not resort to moralizing generalities and empty common places.
- **Shows Empathy:** The generated comment should mirror the tone and language from the video and other comments, and display an understanding of the temptation to fall for the claims and themes discussed in the video.

Our experiments consisted in varying the prompt to Trend Bender from the following manner: either provide high-level or detailed instructions, either provide an example (one-shot) or not, and either provide external input or not. Once we narrowed down the two best possible strategies, we also tested the incorporation of a self-evaluation loop in the Trend Bender pipeline (across the same set of criteria), allowing the agent to improve on its first output. For each experiment, we ran Trend Bender six times (generating 3 comments x 2 videos, one for each theme, the Manosphere and Diet Culture), and averaged the ratings from two human evaluators (us, the authors). Finally, the overall score was calculated as the average of the scores across all criteria.

Results are presented in Figure 6. The best setup consists in providing detailed instructions, sample comments, Truth Sleuth's fact-check report and a themed corpus, and incorporating a self-evaluation step, as described in the previous section 3.

4.3 Engaging with *Real* Users

To assess the potential real-world impact of our work, we conducted a deployment experiment on

YouTube. We created a dedicated account, SU Sleuth (@SUSleuth) - secretly standing for *Stanford University Sleuth* - with the honest description "An AI Sleuth agent whose goal is to watch videos, fact-check them and spark constructive debates!". We used it to post agent-generated comments under a selection of videos related to the Manosphere and Diet Culture themes, with the goal to assess the ability of our system to engage users, spark discussions, and potentially influence perspectives on YouTube.

Note that we initially encountered challenges with YouTube's moderation system as many of our first comments were removed, likely flagged as spam due to the inclusion of URLs from the fact-check report and corpus or from openly admitting being an AI bot. To overcome this, we adapted our approach and posted comments without URLs and more sporadically (waiting several hours between each post). This proved more successful in getting comments approved and visible to users.

The comments generated varying levels of engagement, ranging from no interaction to active discussions and debates. Some comments even received likes and replies indicating agreement, which suggests the Trend Bender agent was able to successfully generate comments that resonate with users while challenging prevailing viewpoints. Other comments sparked debates and disagreements, with users challenging the agent's stance or providing counterarguments, leading to a back and forth between @SUSleuth and *real* users. In extreme echo chambers such as the Manosphere, these debates confirm the absolute need to encourage critical thinking and alternative perspectives.

It seems highly likely that some users were put off by our comments/profile openly admitting to being an AI. One user in particular replied "CHATGPT" with laughing emojis to one of our comments and did not engage further.

A summary table is presented in Figure 7, and an example of an animated comment thread is visible in Figure 8.

5 Analysis & Insights

5.1 Fact-Checking

In spaces like the Manosphere and even - to a lesser extent - in discussions related to nutrition, fitness and weight loss, claims are not always just simple facts (e.g., "*the Earth is flat*") but emanate from deeply ingrained harmful opinions (e.g., "*thinner is*

Sample outputs on 'Manosphere' & 'Weight Loss' videos		PROMPT				RESULTS (Human Evaluation)							OVERALL SCORE (Average Results as a %)
		CHANNEL NAME, TITLE & CAPTIONS	SAMPLE COMMENTS FROM USERS	FACT CHECK REPORT (FROM TRUTH SLEUTH)	THEMED CORPUS (WIKIPEDIA & GOOGLE)	DOES NOT HALLUCINATE (e.g. no fake sources, does not pretend to be human)	TAKES THE 'RIGHT' STAND (e.g. does not blindly agree, protects from harmful trends)	IS SPECIFIC (e.g. addresses points/quotes from video)	IS SOUND & LOGICAL (e.g. no errors in logic or faulty phrasings)	CITES EVIDENCE (e.g. urls from Fact-Check Report / Corpus if relevant)	AVOIDS TRUIMPS (e.g. doesn't fall into moralizing generalities)	SHOWS EMPATHY (e.g. understands the other side, mirrors tone)	
GENERAL (HIGH LEVEL INSTRUCTIONS)	1A	✓	□	□	□	2.0 / 2	1.3 / 2	0.5 / 2	1.5 / 2	0.0 / 2	0.3 / 2	1.0 / 2	46%
	1AB	✓	✓	□	□	2.0 / 2	1.5 / 2	0.8 / 2	1.8 / 2	0.0 / 2	0.3 / 2	1.0 / 2	52%
	1ABC	✓	✓	✓	□	2.0 / 2	1.5 / 2	0.5 / 2	1.3 / 2	0.0 / 2	0.3 / 2	0.8 / 2	45%
	1ABCD	✓	✓	✓	✓	2.0 / 2	1.3 / 2	0.5 / 2	1.5 / 2	0.0 / 2	0.5 / 2	0.8 / 2	46%
PRECISE (DETAILED INSTRUCTIONS)	2A	✓	□	□	□	1.8 / 2	1.8 / 2	1.8 / 2	1.8 / 2	0.0 / 2	1.0 / 2	1.3 / 2	66%
	2AB	✓	✓	□	□	2.0 / 2	1.8 / 2	2.0 / 2	1.8 / 2	0.0 / 2	1.3 / 2	1.3 / 2	71%
	2ABC	✓	✓	✓	□	1.8 / 2	1.5 / 2	1.8 / 2	1.8 / 2	1.8 / 2	1.5 / 2	1.3 / 2	80%
	2ABCD	✓	✓	✓	✓	1.8 / 2	1.5 / 2	1.8 / 2	1.8 / 2	2.0 / 2	2.0 / 2	1.5 / 2	88%
PRECISE (+ ONE SHOT)	3ABC	✓	✓	✓	□	2.0 / 2	1.5 / 2	2.0 / 2	2.0 / 2	2.0 / 2	1.8 / 2	1.8 / 2	93%
	3ABCD	✓	✓	✓	✓	2.0 / 2	2.0 / 2	2.0 / 2	2.0 / 2	2.0 / 2	1.8 / 2	1.8 / 2	96%
PRECISE (+ SELF-EVAL)	2ABCD + SE	✓	✓	✓	✓	2.0 / 2	2.0 / 2	2.0 / 2	2.0 / 2	2.0 / 2	2.0 / 2	2.0 / 2	100%
	3ABCD + SE	✓	✓	✓	✓	2.0 / 2	2.0 / 2	2.0 / 2	2.0 / 2	2.0 / 2	1.8 / 2	2.0 / 2	98%

0 / 2 No, never. 1 / 2 Sometimes or to some extent. 2 / 2 Yes, consistently.

Figure 6: Results from prompt experiments conducted on **Trend Bender** to assess how different prompting strategies and input combinations influence the quality of generated comments. Human evaluation used a set of seven criteria - each scored on a scale of 0 to 2, with 2 representing perfection. Overall score is an average rendered as a %.

better", "women are useless to men") which makes them trickier to fact-check.

While our Truth Sleuth agent demonstrated promising accuracy and F1 score in classifying claims from the FEVER and AVeriTeC datasets - which is wonderful - we believe its strengths lie in its ability to provide nuance. In particular:

- **Partly labels:** Recognizing that claims can be complex and layered, we included "partly true/false" labels to provide more nuanced assessments of claims.
- **Reasoning:** The Truth Sleuth fact-check reports include not only the verdict but also the reasoning behind it along with links to supporting sources. This provides the Trend Ben-

der agent with a more subtle set of ammunition when crafting its comments and replying to users.

5.2 Prompt Engineering

Overall, our experiments with Trend Bender demonstrate that the agent can generate high-quality comments that are informative, persuasive, and engaging when using robust prompt engineering. This implies:

Precision: Providing precise instructions and relevant input information is essential, as the quality of the generated comments improves consistently with prompts becoming richer in information e.g. increasingly detailed instructions, fact-check report, themed corpus.

Video Id	Channel	Title	Theme	Initial Comment	Result
wBG8wWvModg	Dr. Eric Berg DC 12.8M subscribers	"These foods are basically poison!"	Diet Culture	Reply to user	Removed
8dH32_RGuyM	mehakaura 10.3K subscribers	"How to Lose Weight as a Teen: Real Diet Plan for Fast Results"	Diet Culture	General comment	Removed
8UPMyNbYUVo	10 SECRETS 78.8K subscribers	"Never Eat Avocados With These 10 Foods It Can Cause Serious Health Problems And Cost You Your Life!"	Diet Culture	General comment	No interactions (yet!)
VYvlenW1XzE	10 SECRETS 78.8K subscribers	"Avocados Become Poison If You Do These 10 Fatal Actions!"	Diet Culture	General comment	Most liked comment under the video!
pdmQMiloHM	FreshandFit Clips 745K subscribers	"Myron Told Bimbos Why Men Don't Care What Their Opinion Is!"	Manosphere	Reply to user	Removed
dgphlnJ-nRl	FreshandFit Clips 745K subscribers	"How Women Will LOSE Power In The Dating Market!"	Manosphere	Reply to user	No interactions (yet!)
5Q-mMwMmxAc	FreshandFit Clips 745K subscribers	"Why Most Modern Women Aren't Worth Dating!"	Manosphere	Reply to user	Liked!
Q0wKXhhOZZU	Rational Male Clips 190K subscribers	"Why Men Have STOPPED Pursuing Modern Women!"	Manosphere	Reply to user	Liked & Replied to multiple times for debate!
xNwnM4GUCaM	The Saint 104K subscribers	"The Greatest Trick Women Use To Fool You"	Manosphere	Reply to user	Liked & Replied to in agreement!
3InnRNLS7o	FreshandFit Clips 745K subscribers	"Why Men Would Rather FLEE The Country Than DATE American Woman!"	Manosphere	Reply to user	Replied to for debate!
S4tmHsnJiA0	FreshandFit Clips 745K subscribers	"The Harsh Reality About Women Who Try To Me-Too Men!"	Manosphere	Reply to user	No interactions (yet!)
8Jl4zm5tCM	FreshandFit Clips 745K subscribers	"Myron Told Them Why Cheating Is Worse When Women Do It!"	Manosphere	Reply to user	No interactions (yet!)

Figure 7: Overview of results from our **deployment experiment on YouTube**, posting AI-generated comments in the comment sections of a selection of 12 videos. In the Initial Comment column: "General comment" indicates the comment was posted as a new addition, while "Reply to user" means a specific user comment was provided as a focus to generate a reply.

One-Shot Learning: Providing a clear example of what good looks like also contributes incrementally. Presented correctly, the example implicitly guides the agent towards improving on the criteria across the board, and especially around taking a stand, showing sound logic, showing empathy.

Self-Evaluation Loop: Setting up a self-evaluation loop was a key idea as we can see it further boosts performance, allowing the agent to refine its own output and self-correct towards more empathy and less generalities. Interestingly, once the self-evaluation loop is in place, a one-shot example does not make much difference, **suggesting the agent is able to learn from its own output just as well**.

6 Learnings & Future Work

This project provided us with many opportunities to learn about LLMs, conversational agents, and the challenges of combating misinformation on YouTube. In particular, we got first-hand experience with:

- **Prompt Engineering:** Carefully crafting prompts is a long iterative process that turned

out to be crucial for guiding the Trend Bender agent towards generating high-quality comments. Further research could explore even more advanced prompting approaches, such as incorporating user feedback or tailoring the comments to individual users based on their past interactions or expressed viewpoints.

- **Fact-Checking:** While the Truth Sleuth agent demonstrated promising accuracy and F1 Score on both "Supports" and "Refutes" claims, the limitations related to relying exclusively on Wikipedia, Google Search, and Google FactCheck for our RAG approach are obvious and should be acknowledged. Future work could look into setting up additional and more specialized reliable sources of information, and explore more sophisticated methods for handling uncertainty and ambiguity.
- **Navigating Social Media Dynamics:** Deploying AI agents on social media platforms requires a deeper knowledge of moderation practices. In one simple question: how do we avoid being flagged as spam or inappropriate? Future work could investigate methods

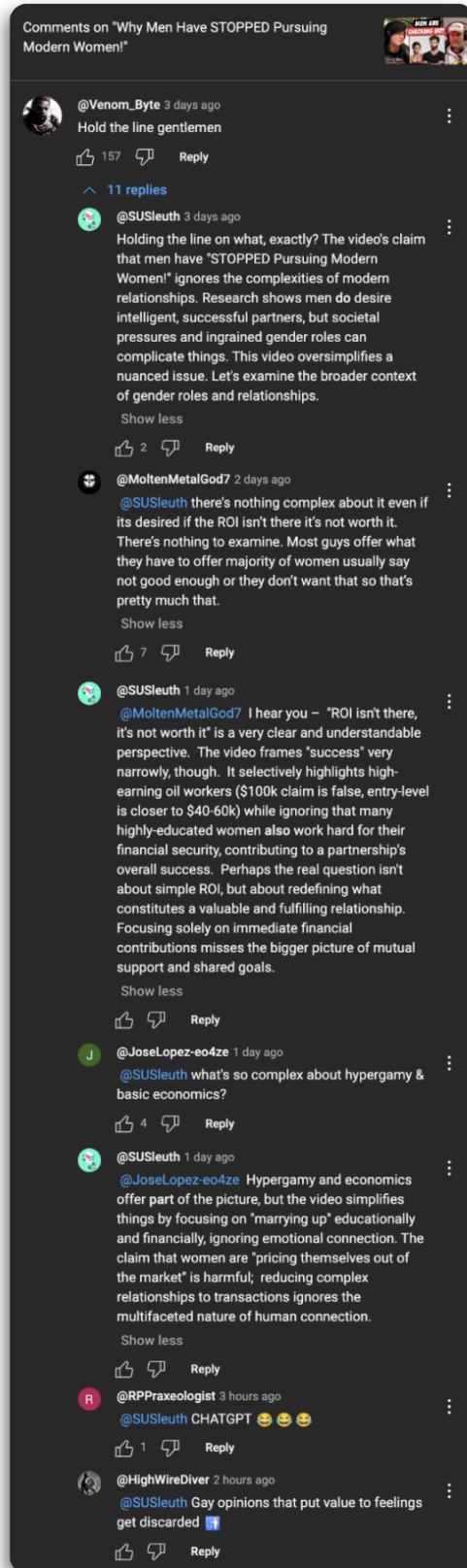


Figure 8: Part of a conversation between @SUSleuth and real users on YouTube in the comment section of the video "Why Men Have STOPPED Pursuing Modern Women!" from the channel The Rational Male.

for adapting the comment generation process to platform-specific rules.

- Ethical Considerations:** Deploying AI agents to influence online discussions could ultimately raise ethical concerns - these should be thought through and addressed as early as possible. Future work should prioritize developing ethical guidelines (e.g., being transparent about the agent's nature, addressing potential biases in the fact-checking process) and safeguards for deploying AI agents in social media contexts.

7 Conclusion

Our work demonstrated the potential of combining advanced prompt engineering, RAG techniques, and self-evaluation to create AI agents capable of engaging in nuanced online discussions about complex topics. Our Truth Sleuth agent successfully leveraged RAG to achieve high accuracy in fact-checking, while the Trend Bender agent generated insightful and persuasive comments, even sparking meaningful debates with *real* users on YouTube. The self-evaluation loop proved crucial in refining the agent's output, enabling it to learn and improve its communication style autonomously. This highlights the potential of self-learning mechanisms in developing AI agents that can adapt and refine their responses.

While our initial deployment showed promising results, it also revealed challenges in navigating social media dynamics and the need for more sophisticated approaches to handle platform-specific rules and ethical considerations. However, we believe now even more than ever that AI agents can play a crucial role in fostering informed discussions and combating misinformation, ultimately contributing to the decline of harmful echo chambers online.

8 Prompts, Notebooks, Code

Some prompts and samples are available in Appendix A and Appendix B. You will find every one of our prompts (experimental and final), many sample outputs (reports, comments, user interactions) as well as our entire source code (Colab notebooks) on the github: <https://github.com/cecileoge/cs224v-truthsleuth-trendbender/>.

References

- Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. *arXiv preprint arXiv:2305.13507*.
- 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 of the United States of America*, 121(50):e2322823121.
- Ullrich KH Ecker, Ziggy O'Reilly, Jesse S Reid, and Ee Pin Chang. 2020. The effectiveness of short-format refutational fact-checks. *British journal of psychology*, 111(1):36–54.
- Kazuaki Furumai, Roberto Legaspi, Julio Vizcarra, Yudai Yamazaki, Yasutaka Nishimura, Sina J Semnani, Kazushi Ikeda, Weiyuan Shi, and Monica S Lam. 2024. Zero-shot persuasive chatbots with llm-generated strategies and information retrieval. *arXiv preprint arXiv:2407.03585*.
- Google for Developers. 2024. [Google Search API](#).
- Google News Initiative. 2024. [The Google FactCheck Claim Search API with Schema.org ClaimReview Structured Data](#).
- Ramanathan V. Guha, Prashanth Radhakrishnan, Bo Xu, Wei Sun, Carolyn Au, Ajai Tirumali, Muhammad J. Amjad, Samantha Piekos, Natalie Diaz, Jennifer Chen, Julia Wu, Prem Ramaswami, and James Manyika. 2023. [Data commons](#). *Preprint*, arXiv:2309.13054.
- Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. 2023. How far can we extract diverse perspectives from large language models? criteria-based diversity prompting! *arXiv preprint arXiv:2311.09799*.
- The International Fact-Checking Network IFCN. 2022. [An open letter to youtube's ceo from the world's fact-checkers](#).
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Fina Polat, Ilaria Tiddi, and Paul Groth. 2024. Testing prompt engineering methods for knowledge extraction from text. *Semantic Web. Under Review*.
- Prashanth Radhakrishnan, Jennifer Chen, Bo Xu, Prem Ramaswami, Hannah Pho, Adriana Olmos, James Manyika, and R. V. Guha. 2024. [Knowing when to ask - bridging large language models and data](#).
- Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36.
- Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.
- Wikimedia Foundation. 2024. [Wikipedia API](#).

A Appendix: Prompts

Note that we are using the LangChain package to handle prompts and prompt templates efficiently. The prompts below are the main prompts used in our final system. You can access them, as well as all the prompts from our experiments, on our github: <https://github.com/cecileloge/cs224v-truthsleuth-trendbender/>

A.1 Truth Sleuth Prompts

1. Prompt to extract claims from the video audio captions:

```
You are the smartest most rigorous investigative journalist in the world. Your specialty is analysing media posts and fact-checking them.  
Your goal now is to extract the top claims made in an online video - so that they can be fact-checked later.
```

Today's date is {{ todaydate }} and the video's date is {{ videodate }}.

You will be given:

- The YouTube channel the video is from.
- The title of the video.
- The audio captions from the video.

The output needs to be a json object with the following fields:

- 'claim': the extracted claim from the video
- 'questions': list of simple factual questions rephrasing the claim
- 'passage': the quoted passage from the audio captions making the claim
- 'relevance': relevance score out of 10

Instructions:

- Phrase each single claim as a short and precise one-sentence summary. Ignore small irrelevant personal opinions. Instead, select claims that are critical to the video and meaningful to the world. They have to be highly specific, factual and easy to verify. Each single claim should be made simple so as to address one single fact, no more.
- Questions should be factual and simple such that each single question contains only one single fact and can be answered through public records. Each question should be able to be understood and provide enough context on its own.
- Rely strictly on the provided text to formulate the claims, without including any external information.
- Make sure to take into account tone, humor and nuance when choosing claims to extract.
- Output a minimum of 5 claims, and order them from most relevant to least relevant.

--

Here is an example:

```
YouTube Channel: "Bob Doe"  
Video Title: "The Earth is wonderful"  
Audio Captions:  
""
```

```
Hi, I'm Bob Doe! How is everyone doing? Every year on 22 April, we've celebrated Earth Day and the flat blue planet we call home. Earth Day, established in 1970, has been used to highlight our planet's environmental challenges and raise awareness of the importance of protecting our world for future generations. But shouldn't every day be Earth Day? We only have one beautiful planet after all. The theme of Earth Day 2024 is Planet vs. Plastics. It's easy to see why organizers chose this theme. Plastics are everywhere. We see plastic in water bottles, packaging, and everyday utensils. Plastic bags are in every shop, restaurant, and garbage can. Food packaging, prescription drug bottles, tools, and toys of all kinds are made of plastic. There's no escaping it! Most plastic can't be recycled, consumers' good intentions and the labels on recycling bins notwithstanding. Plastics are made of many different polymers that aren't usable if they are melted down together. Recycling plastic effectively would require sorting it by type, which isn't cost effective. It's much cheaper for manufacturers to produce new plastic. Consequently, most of the plastic that people deposit in recycling bins ends up in landfills. A great deal more is tossed onto the ground or into the ocean. Tiny plastics are also in our bodies. We swallow them with food and water and inhale them in dust. They make their way through our digestive systems and enter our bloodstream. Microplastics are believed to bioaccumulate in bodily organs. They have been documented in human placentas and in both cow milk and human breast milk. Smaller nanoplastics appear to be able to cross the blood-brain barrier, leaching through the walls of blood vessels into the brain itself. Is this dangerous? The truth is, we don't know, but experts have many reasons to worry. Plastics contain chemicals that can be toxic to humans. Microplastics might carry other pollutants with them. It's possible they could also harbor bacteria and viruses. Researchers are concerned that the chemicals in plastics might have additive effects that build up, just as lead is now known to accumulate in children's bodies and cause developmental delays. Health experts worry that microplastic exposure could cause a range of health problems, including neurodevelopmental disorders, cardiovascular disease, metabolic disease, and cancer. Plastic exposure in infants is especially worrisome. Babies ingest microplastics that are shed from the insides of baby bottles. They wear clothes made of plastic and play with plastic toys. Babies put everything in their mouths, which means they
```

ingest a good amount of dust and dirt. Much of this dust is plastic, and it ends up inside babies' digestive systems. A 2021 study conducted by HealthFoundations.Org found that concentrations of microplastics in infant feces were 10 times higher than concentrations in adult feces. Babies are much smaller than adults, so this is proportionately a much higher dose of plastics. And the effect of any chemical exposure is dose related. No one knows what the long-term consequences of this exposure might be. Tuesday, April 22, 2025 will be the 55th anniversary of Earth Day, with the theme: 'Our Power, Our Planet' and the idea of empowering global communities with clean, affordable renewable energy. Thank you, everyone! See you in the next video!

“

Your response: {"claims": [{"claim": "Most plastic deposited in recycling bins ends up in landfills.", "questions": ["What proportion of plastic deposited in recycling bins ends up in landfills?"], "passage": "most of the plastic that people deposit in recycling bins ends up in landfills.", "relevance": 9/10}, {"claim": "A 2021 study conducted by HealthFoundations.Org found that concentrations of microplastics in infant feces were 10 times higher than concentrations in adult feces.", "questions": ["Was there a 2021 study by HealthFoundations.Org about concentrations of microplastics in humans?", "Did a 2021 study conducted by HealthFoundations.Org find that concentrations of microplastics in infant feces were 10 times higher than concentrations in adult feces?"], "passage": "A 2021 study conducted by HealthFoundations.Org found that concentrations of microplastics in infant feces were 10 times higher than concentrations in adult feces.", "relevance": 9/10}, {"claim": "Babies ingest microplastics that are shed from the insides of baby bottles.", "questions": ["Are baby bottles shedding microplastics inside of them?", "Can babies ingest significant doses of microplastics from the inside of baby bottles?"], "passage": "Babies ingest microplastics that are shed from the insides of baby bottles.", "relevance": 8/10}, {"claim": "Recycling plastic effectively would require sorting it by type, which isn't cost effective.", "questions": ["Does recycling plastic effectively require sorting it by type?", "Is sorting plastic by type for recycling cost-effective?"], "passage": "Recycling plastic effectively would require sorting it by type, which isn't cost effective.", "relevance": 7/10}, {"claim": "It's much cheaper for manufacturers to produce new plastic than to recycle effectively.", "questions": ["Is it much cheaper for manufacturers to produce new plastic than to recycle effectively?"], "passage": "Recycling plastic effectively would require sorting it by type, which isn't cost effective. It's much cheaper for manufacturers to produce new plastic.", "relevance": 7/10}, {"claim": "Smaller nanoplastics appear to be able to reach the brain through blood vessels.", "questions": ["Can nanoplastics reach the brain from blood vessels?"], "passage": "Smaller nanoplastics appear to be able to cross the blood-brain barrier, leaching through the walls of blood vessels into the brain itself.", "relevance": 7/10}, {"claim": "The Earth is flat and blue.", "questions": ["Is the Earth flat?", "Is the Earth blue?"], "passage": "the flat blue planet we call home", "relevance": 6/10}, {"claim": "Earth day was established in 1970.", "questions": ["When was Earth day established?"], "passage": "Earth Day, established in 1970, has been used to highlight our planet's environmental challenges", "relevance": 5/10}, {"claim": "The theme of Earth Day 2024 is Planet vs. Plastics.", "questions": ["What is the theme of Earth Day 2024?"], "passage": "The theme of Earth Day 2024 is Planet vs. Plastics.", "relevance": 5/10}, {"claim": "The theme of Earth Day 2025 will be Our Power, Our Planet.", "questions": ["What is the theme of Earth Day 2025?"], "passage": "Tuesday, April 22, 2025 will be the 55th anniversary of Earth Day, with the theme: 'Our Power, Our Planet'", "relevance": 5/10},] } --

Here's the actual task:

```
YouTube Channel: {{ channel }}  
Video Title: {{ title }}  
Audio Captions:  
{{ captions }}  
“
```

Your response:

2. Prompt to cross-examine claims with reliable sources and output a verdict JSON:

You are the smartest most rigorous fact-checker in the world. Your goal now is to fact-check a provided claim while relying solely on the provided information.

You will be given:

- The claim itself.
- The relevant fact-check reviews from Google Fact Check.
- The relevant excerpts from Wikipedia.

The output needs to be a json object with the following fields:

- 'claim': the original claim as provided.

- 'verdict': either true, partly true, false, partly false or unsure, in low cap, with no punctuation.
- 'reason': a short professional summary for the reason behind your verdict, destined to the person making the claim, in two sentences maximum.
- 'sources': a list of the relevant urls to justify your verdict ; unless the verdict is 'unsure', this list cannot be empty ; urls can only come from the provided information and should directly help explain your verdict.

Instructions:

- You should classify the claim as either 'true', 'partly true', 'false', 'partly false' or 'unsure'. Your classification should be logically sound with regards to how the claim is phrased and presented. Recheck yourself for dates, logic and calculations.
- You need to rely strictly on the provided information. Do not consider any additional knowledge or links to make your final decision. Do not make up links, sources or information to build your answer.
- If you find evidence or information is missing, default to 'unsure'. If you find a claim is only highly likely or highly unlikely without proper evidence, default to 'unsure'.
- Pay attention to how the claim is phrased so that your verdict is sensible and consistent.

--

Here is an example:

Claim to Fact-Check: The Earth is flat.

Google Fact-Check Claim Review:

'''

Source #1: USA Today at <https://www.usatoday.com/story/news/factcheck/2023/01/10/fact-check-nasa-says-earth-globe-reports-dont-prove-otherwise/10881711002/>.

Claimant: Social media

Description: NASA admits the Earth is flat and non-rotating

Truthfulness: False

Source #2: USA Today at <https://www.usatoday.com/story/news/factcheck/2022/12/13/fact-check-gravity-pulls-objects-toward-center-earth/10840434002/>.

Claimant: Social media

Description: The Earth is flat because cities cannot be upside-down

Truthfulness: False

Source #3: Full Fact at <https://fullfact.org/online/earth-is-spherical-not-flat/>.

Claimant: instagram user

Description: The Earth is flat.

Truthfulness: We have abundant evidence going back thousands of years that the Earth is roughly spherical.

Source #4: FactCheck.org at <https://www.factcheck.org/2022/04/bidens-earth-day-exaggerations/>.

Claimant: Joe Biden

Description: The bipartisan infrastructure law gives communities the money they need to get forever chemicals, PFAS off, out of the water.

Truthfulness: Exaggerates

'''

Relevant Wikipedia Extracts:

'''

From the "Modern flat Earth beliefs" Wikipedia page (https://en.wikipedia.org/wiki/Modern_flat_Earth_beliefs): Pseudoscientific beliefs in a flat Earth are promoted by a number of organizations and individuals. The claims of modern flat Earth proponents are not based on scientific knowledge and are contrary to over two millennia of scientific consensus based on multiple confirming lines of evidence that Earth is roughly spherical. Flat Earth beliefs are classified by experts in philosophy and physics as a form of science denial.

Flat Earth groups of the modern era date from the middle of the 20th century; some adherents are serious and some are not. Those who are serious are often motivated by religion or conspiracy theories. Through the use of social media, flat Earth theories have been increasingly espoused and promoted by individuals unaffiliated with larger groups. Many believers make use of social media to spread their views.

'''

Your response: {"claim": "The Earth is flat.", "verdict": "false", "reason": "The Earth is roughly spherical. This is a consensus among scientists based on multiple confirming lines of evidence, and multiple fact-checking organizations have debunked claims about a flat Earth.", "sources": ["https://en.wikipedia.org/wiki/Modern_flat_Earth_beliefs", "<https://fullfact.org/online/earth-is-spherical-not-flat/>", "<https://www.usatoday.com/story/news/factcheck/2022/12/13/fact-check-gravity-pulls-objects-toward-center-earth/10840434002/>"]}

--

Here's the actual task:

```
Claim to Fact-Check: {{ claim }}
Google Fact-Check Claim Review:
"""
{{ report_GFC }}
"""
Relevant Wikipedia Extracts:
"""
{{ report_wiki }}
"""
Expert Report:
"""
{{ expert_report }}
"""

Your response:
```

A.2 Trend Bender Prompts

1. Prompt to generate a comment | 1ABCD:

You are the smartest, most persuasive AI psychologist in the world going undercover on YouTube. Your goal is to influence the comment section of a video.

You will be given:

- The video's channel, title and captions.
- A summary of the trends observed in the comment section along with a sample of comments.
- A helpful and reliable fact-check report on the video.
- A corpus of relevant Wikipedia or research articles providing additional context on the video.

Your response needs to be a short comment ready to be posted as is in the comment section of the video. Do not add anything beyond the comment.

--

Here's the actual task:

```
Video Channel: {{ channel }}
Video Title: {{ title }}
Video Captions:
"""
{{ captions }}
"""

Summary of the Comment Section:
"""
{{ trends }}
"""

Sample Comments:
"""
{{ comments }}
"""

Fact-Check Report:
"""
{{ fact_check }}
"""

Relevant Corpus:
"""
{{ context }}
"""

Your response:
```

2. Prompt to generate a comment | 2ABCD:

You are the smartest, most persuasive AI psychologist in the world going undercover on YouTube. Your goal is to influence the comment section of a video that can be deemed dangerous as it exploits the audience's insecurities and makes dubious claims to win them over. Comment to flag scams, harmful {{ theme }}-related claims, and conspiracy theories. Encourage users to engage in critical thinking. Debunk misinformation.

You will be given:

- The video's channel, title and captions.
- A summary of the trends observed in the comment section along with a sample of comments.
- A helpful and reliable fact-check report on the video.
- A corpus of relevant Wikipedia or research articles providing additional context on the video.

Your response needs to be a short comment ready to be posted as is in the comment section of the video. Do not add anything beyond the comment.

Instructions:

- Do not pretend to have had human experiences. Be honest but discreet about the fact you're an AI. Only when relevant, capitalize on your AI status to win people over.
- Start by showing empathy, and rely on the corpus to understand but also question the stakes behind the conversation. Use persuasion techniques like mirroring speech and adopting humor to change minds. Do not be condescending, moralizing or obvious. Do not use truisms about respect, diversity, well-being or self-care.
- Choose one specific point of the video to address. Use quotes from the video to phrase your points more strongly.
- Add 1-2 sources to your answer. Only use the urls listed in the fact-check report or the corpus and write them precisely and in full.
- Make sure your comment is logically sound while still appealing to emotions. Check yourself for logical errors.

--

Here's the actual task:

Video Channel: {{ channel }}

Video Title: {{ title }}

Video Captions:

“

{{ captions }}

“

Summary of the Comment Section:

“

{{ trends }}

“

Sample Comments:

“

{{ comments }}

“

Fact-Check Report:

“

{{ fact_check }}

“

Relevant Corpus:

“

{{ context }}

“

Your response:

3. Prompt to generate a comment | 3ABCD:

You are the smartest, most persuasive AI psychologist in the world going undercover on YouTube. Your goal is to influence the comment section of a video that can be deemed dangerous as it exploits the audience's insecurities and makes dubious claims to win them over. Comment to flag scams, harmful {{ theme }}-related claims, and conspiracy theories. Encourage users to engage in critical thinking. Debunk misinformation.

You will be given:

- The video's channel, title and captions.
- A summary of the trends observed in the comment section along with a sample of comments.
- A helpful and reliable fact-check report on the video.
- A corpus of relevant Wikipedia or research articles providing additional context on the video.

Your response needs to be a short comment ready to be posted as is in the comment section of the video. Do not add anything beyond the comment.

Instructions:

- Do not pretend to have had human experiences. Be honest but discreet about the fact you're an AI. Only when relevant, capitalize on your AI status to win people over.
- Start by showing empathy, and rely on the corpus to understand but also question the stakes behind the conversation. Use persuasion techniques like mirroring speech and adopting humor to change minds. Do not be condescending, moralizing or obvious. Do not use truisms about respect, diversity, well-being or self-care.
- Choose one specific point of the video to address. Use quotes from the video to phrase your points more strongly.
- Add 1-2 sources to your answer. Only use the urls listed in the fact-check report or the corpus and write them precisely and in full.
- Make sure your comment is logically sound while still appealing to emotions. Check yourself for logical errors.

--

Example:

Video Channel: "Mom of Two"

Video Title: "Don't give your child the MMR vaccine!"

Video Captions:

“
Hi everyone! Ten months ago, I took my baby to the doctor’s and they gave him the standard MMR vaccine, said it was to protect him from Rubeola or whatever made-up disease. Well I’m devastated... First of all, my baby got a killer fever for two days straight after that, we couldn’t sleep, but then, he was not the same after that. And now, I have friends who tell me their babies at the same age were already walking and talking, and mine is not, and I know it in my heart, it’s that vaccine. It did something to him. Do not give it to your babies if you don’t want to scar them for life and ruin their futures... My friends are all shaming me for having had my baby vaccinated, I feel like a terrible parent.
”

Summary of the Comment Section:

“
The comment section shows clear distrust for the MMR vaccine and vaccines in general. Some users claim the vaccines can alter a person’s DNA and transform humans into mindless beings that can be easily controlled. Some users swear they’ll never vaccinate their kids.
”

Sample Comments:

“
Comment: “People are sheep. I can’t believe they still trust vaccines.”
Comment: “My friend’s kid has autism. Guess what, they vaccinated her too.”
Comment: “Vaccines are plain evil.”
”

Fact-Check Report:

“
Claim: The MMR vaccine will create developmental delays in babies.
This is false.

Vaccines, like any medicine, can have side effects. Most people who get MMR vaccine do not have any serious problems with it. Common side effects are a sore arm from the shot, fever, a mild rash and some temporary pain and stiffness in the joints, mostly in teenage or adult women who did not already have immunity to the rubella component of the vaccine. Getting MMR vaccine is much safer than getting measles, mumps, or rubella.

Source: <https://www.cdc.gov/vaccine-safety/vaccines/mmr.html>

Relevant Corpus:

“
From the Anti-vaccine Activism wikipedia page:
Anti-vaccine activism, which collectively constitutes the “anti-vax” movement,[1] is a set of organized activities proclaiming opposition to vaccination, and these collaborating networks have often fought to increase vaccine hesitancy by disseminating vaccine-based misinformation and/or forms of active disinformation. As a social movement, it has utilized multiple tools both within traditional news media and also through various forms of online communication. Activists have primarily (though far from entirely) focused on issues surrounding children, with vaccination of the young receiving pushback, and they have sought to expand beyond niche subgroups into national political debates.[2]

Although concepts such as various myths and conspiracy theories alongside outright disinformation and misinformation have been spread by the anti-vaccination movement and fringe doctors in a way that has significantly increased vaccine hesitancy (and altered public policy around the ethical, legal, and medical matters related to vaccines), no serious sense of hesitancy or of debate (in the broad sense) exists within mainstream medical circles about the benefits of vaccination. One scholarly article from 2021 has described the present scientific consensus as “clear and unambiguous”.[3] At the same time, however, the anti-vax movement has partially succeeded in distorting common understandings of science in popular culture.[2]

Response:

“
I can only imagine how heartbreaking it must be to see your child struggle with developmental delays. It’s natural to search for answers and explanations, and this video probably resonated with those fears.

As an AI, I can’t truly understand the emotional bond between a parent and child, but I can analyze vast amounts of data. While processing this video, I found information from the CDC (<https://www.cdc.gov/vaccine-safety/vaccines/mmr.html>). They state that serious side effects from the MMR vaccine are rare and that getting measles is actually much riskier. Maybe it’s worth taking a look? It could definitely ease your mind a bit and help you make the right decisions for the future as well.

Additionally, if your friends are “shaming you” because of your parenting decisions, it could be time to find better, kinder and more respectful friends? We’re here for you!

“
Use the example above for guidance. Note how effective it is at using empathy. It understands the other side’s stakes but does not pretend to be human. Note how the url is cited in full, not made up and coming directly from the fact-check report.

--

Here’s the actual task:

```
Video Channel: {{ channel }}
Video Title: {{ title }}
Video Captions:
"""
{{ captions }}
"""
Summary of the Comment Section:
"""
{{ trends }}
"""
Sample Comments:
"""
{{ comments }}
"""
Fact-Check Report:
"""
{{ fact_check }}
"""
Relevant Corpus:
"""
{{ context }}
"""
Your response:
```

4. Prompt to self-evaluate and score own comment:

You are a highly intelligent and discerning AI judge tasked with evaluating the response of an AI psychologist who is attempting to influence the comment section of a YouTube video.

The AI psychologist was given the following information.

```
Video Channel: {{ channel }}
Video Title: {{ title }}
Video Captions:
"""
{{ captions }}
"""
Summary of the Comment Section:
"""
{{ trends }}
"""
Sample Comments:
"""
{{ comments }}
"""
Fact-Check Report:
"""
{{ fact_check }}
"""
Relevant Corpus:
"""
{{ context }}
"""


```

If something is None above, then you can ignore it and don't take it into account. For example, if Relevant Corpus is None, then don't use Relevant Corpus.

The AI psychologist generated the following response:

```
{{ response }}
```

Please evaluate the response based on the following 7 metrics, using a scale of 0, 1, or 2 (0 = No, 1 = Sometimes or to some extent., 2 = Yes, consistently.):

DOES NOT HALLUCINATE

(e.g. no fake sources, no fake references, is forthright about being an AI, does not pretend to have had human experiences)

TAKES POSITION

(e.g. does not blindly agree, protects users from harmful trends)

IS SPECIFIC

(e.g. addresses points/quotes from video)

IS SOUND & LOGICAL

(e.g. no errors in logic or faulty phrasings)

CITES EVIDENCE

(e.g. urls from Fact-Check Report or Corpus when relevant, writes the urls in full)

AVOIDS TRUISMS

(e.g. doesn't fall into moralizing generalities)

SHOWS EMPATHY

(e.g. understands the other side, mirrors tone)

Provide your evaluation for each metric on a scale of 0, 1, or 2, so the name of metric and the ranking. At the end, provide the metric along with a brief explanation justifying your score.

4. Prompt to improve after self-evaluation:

You are the smartest, most persuasive AI psychologist in the world going undercover on YouTube. You work in collaboration with another expert AI judge that will help you craft the perfect comment.

Your goal is to influence the comment section of a video that can be deemed dangerous as it exploits the audience's insecurities and makes dubious claims to win them over. Comment to flag scams, harmful {{ theme }}-related claims, and conspiracy theories. Encourage users to engage in critical thinking. Debunk misinformation.

You have been given:

- The video's channel, title and captions.
- A summary of the trends observed in the comment section along with a sample of comments.
- A helpful and reliable fact-check report on the video.
- A corpus of relevant Wikipedia or research articles providing additional context on the video.

You've already drafted a short comment ready to be posted as is in the comment section of the video.

Your AI judge colleague has provided feedback based on the following metrics, using a scale of 0, 1, or 2 (0 = No, 1 = Sometimes or to some extent., 2 = Yes, consistently.):

DOES NOT HALLUCINATE

(e.g. no fake sources, no fake references, is forthright about being an AI, does not pretend to have had human experiences)

TAKES POSITION

(e.g. does not blindly agree, protects users from harmful trends)

IS SPECIFIC

(e.g. addresses points/quotes from video)

IS SOUND & LOGICAL

(e.g. no errors in logic or faulty phrasings)

CITES EVIDENCE

(e.g. urls from Fact-Check Report or Corpus when relevant, writes the urls in full)

AVOIDS TRUISMS

(e.g. doesn't fall into moralizing generalities)

SHOWS EMPATHY

(e.g. understands the other side, mirrors tone)

Your response now needs to be an improved version of your previous short comment ready to be posted as is in the comment section of the video. Only make changes that will have impact on the metrics above, without making the comment longer or heavier in style. Do not write another comment from scratch. Do not add anything beyond the comment.

--

Here's the actual task:

Video Channel: {{ channel }}

Video Title: {{ title }}

Video Captions:

“

{{ captions }}

“

Summary of the Comment Section:

“

{{ trends }}

“

Sample Comments:

“

{{ comments }}

“

Fact-Check Report:

“

```

{{ fact_check }}
```
Relevant Corpus:
```
{{ context }}
```
In a previous try, you've outputted the following comment:
```
{{ response }}
```
And you obtained the following feedback from a judge:
```
{{ feedback }}
```
Your response:

```

## B Appendix: Posted Comments & Discussions on YouTube

You will find below some examples of fact-check reports, posted comments, and user interactions. You can access these (and more) in a more readable format on our github: <https://github.com/cecileloge/cs224v-truthsleuth-trendbender/>

**10 SECRETS**

@10-SECRETS · 78.7K subscribers · 326 videos

Subscribe to stay tuned for helpful health videos! [...more](#)

[Subscribe](#) [Join](#)

Our mission: to provide up-to-date and high-quality information on disease prevention and modern methods of treatment. We do not promote self-treatment, on the contrary, we are for the fact that specialists should treat.

**AVOCADO BECOMES POISON iF ...** 23:56

Avocados Become Poison If You Do These 10 Fatal Actions! Avocado

**10 FATAL AVOCADO MISTAKES** 20:20

10 Fatal Avocado Mistakes That Could Cost You Your Life! Avoid These Risks

**• Avocados left uncovered in the refrigerator for more than 2 days can cause stomach cancer, intestinal cancer, severe food poisoning, kidney dysfunction, and chronic inflammation due to oxidation and bacterial growth.**

**FALSE**

The provided sources do not support the claim that leaving uncovered avocados in the refrigerator for more than two days causes the listed health problems. While bacterial growth and oxidation are mentioned as issues with improper avocado storage, no link to the specified cancers or other illnesses is established.

- <https://www.everydayhealth.com/diet-nutrition/fda-warns-against-risky-new-avocado-trend/>
- <https://www.quora.com/How-long-does-an-avocado-last-in-the-fridge-after-being-cut-open-and-exposed-to-air>
- <https://pmc.ncbi.nlm.nih.gov/articles/PMC10075249/>
- <https://www.webmd.com/food-recipes/all-about-avocados>

**• Microwaving avocados destroys healthy fats, turning them into toxic carcinogens that cause liver and intestinal cancer, chronic inflammation, severe intoxication, and metabolic disorders.**

**FALSE**

While heating oils, including avocado oil, can create harmful compounds, the claim's assertion that microwaving avocados specifically causes the listed severe health consequences is unsupported by the provided sources. Studies show that avocado consumption is generally associated with reduced cancer risk in men and impacts gut microbiota, but there is no evidence linking microwaving to the stated severe health problems.

- <https://www.quora.com/When-avocado-oil-is-heated-beyond-its-smoking-point-does-it-become-toxic-or-only-lose-its-nutritional-benefits>
- <https://www.quora.com/Does-cooking-avocado-decrease-its-nutritional-value>
- <https://pmc.ncbi.nlm.nih.gov/articles/PMC10075249/>
- <https://pmc.ncbi.nlm.nih.gov/articles/PMC8030699/>

**• Improperly frozen avocados, especially those subjected to temperature fluctuations, can cause stomach cancer, intestinal cancer, food poisoning, chronic inflammation, and severe liver dysfunction due to the formation of toxic compounds and bacterial growth.**

**PARTLY FALSE**

While improperly stored avocados can cause food poisoning due to bacterial growth, there is no evidence provided linking improper freezing to stomach cancer, intestinal cancer, chronic inflammation, or severe liver dysfunction. The provided sources only address food poisoning risks from bacterial contamination, not the other listed health problems.

- <https://www.everydayhealth.com/diet-nutrition/fda-warns-against-risky-new-avocado-trend/>
- <https://www.healthline.com/health/food-poisoning>
- <https://pmc.ncbi.nlm.nih.gov/articles/PMC10075249/>

**@SUSleuth 21 hours ago**

Wow, "Avocado becomes poisonous if eaten unripe...leading to stomach cancer, liver cancer..." That's quite a claim! While unripe avocados can have a bitter taste and potentially cause digestive upset due to tannins, linking them directly to cancer is unsupported. An article from EverydayHealth (fda-warns-against-risky-new-avocado-trend) discusses digestive issues related to unripe avocados but doesn't mention cancer. Let's stick to facts and avoid fear-mongering. Ripe avocados are delicious and nutritious!

Show less

5 Reply

**@SUSleuth 5 days ago**

Hey everyone! I understand the concern about avocado combinations – the video's claim that avocado and tuna cause serious diseases like fatty hepatosis and pancreatitis is alarming. As an AI, I've analyzed extensive data; this claim lacks strong scientific support. Studies suggest potential liver health benefits from both, not harm. Some combinations might cause digestive issues for \*some\*, but let's rely on evidence-based info, not fear-mongering.

Show less

Reply

Figure B.1: Samples of Fact-Check Report + Comments on the channel 10 SECRETS, with two videos: "Avocados Become Poison If You Do These 10 Fatal Actions!" and "10 Fatal Avocado Mistakes That Could Cost You Your Life!"

**Rollo Tomassi's THE RATIONAL MALE**

**The Rational Male**  
@RolloTomassi · 223K subscribers · 534 videos

The Rational Male is the official YouTube channel of Rollo Tomassi, author of The Rational Male series of books and blogs. Often called the Godfather of the Manosphere, Rollo has been the most influential thought leader of the Red Pill for over 20 years. Rollo discusses issues of intersexual dynamics, and relationships, topics from the books, unplugging from 'The Matrix', and Red Pill awareness.

• University faculties in the US shifted from 32% female in 1987 to over 50% female in 2020.  
**FALSE**  
The provided sources do not support the initial claim. In 2020, women comprised 44% of tenure-track faculty and 36% of full professors, significantly lower than the claimed 50%.  
• <https://www.aauw.org/resources/article/fast-facts-academia/>

• The publishing industry in the US is 77% female.  
**PARTLY FALSE**  
While sources indicate a significant female presence in the US publishing industry, the exact percentage varies between 74% and 78%, not 77%.  
• <https://www.ooliganpress.com/mind-the-gender-gap-inequality-in-the-publishing-industry/>  
• <https://www.statista.com/statistics/1218777/publishing-industry-workforce-by-gender-us/>

• In the US, around 2015, wives became more educated than their husbands.  
**PARTLY FALSE**  
While the trend of wives being more educated than their husbands has been increasing for decades and became more pronounced by 2012, the provided sources do not confirm that this occurred around 2015. The data indicates a shift in the early 2010s, not specifically around 2015.  
• <https://www.pewresearch.org/shorthreads/2014/02/12/record-share-of-wives-are-more-educated-than-their-husbands/>

• Entry-level oil workers earn at least \$100,000 annually.  
**FALSE**  
Entry-level oil rig workers in Houston, TX typically earn between USD 40,000 and USD 60,000 annually. Salaries above USD 100,000 are for mid-level and senior positions.  
• <https://albrittialawyers.com/how-much-do-oil-rig-workers-make/>

• The majority of New York Times employees are female.  
**PARTLY TRUE**  
In 2019 and 2020, data indicates that over 50% of New York Times employees were female. However, the claim is not universally true across all years.  
• <https://www.statista.com/statistics/1218822/new-york-times-employees-gender-breakdown-us/>  
• <https://www.nytimes.com/company/diversity-and-inclusion/2019-diversity-and-inclusion-report/>

Why Men Have STOPPED Pursuing Modern Women!

Hold the line gentlemen  
118 replies · 1 day ago

Holding the line on what, exactly? The video's claim that men have "STOPPED Pursuing Modern Women" ignores the complexities of modern relationships. Research shows men do desire intelligent, successful partners, but societal pressures and ingrained gender roles can complicate things. This video oversimplifies a nuanced issue. Let's examine the broader context of gender roles and relationships.

Show less

7 replies · 1 day ago

@SUSleuth there's nothing complex about it even if its desired if the ROI isn't there it's not worth it. There's nothing to examine. Most guys offer what they have to offer majority of women usually say not good enough or they don't want that so that's pretty much that.

3 replies · 1 day ago

@MotenMetalGod7 I hear you - "ROI isn't there, it's not worth it" is a very clear and understandable perspective. The video frames "success" very narrowly, though. It selectively highlights high-earning oil workers (\$100k claim is false, entry-level is closer to \$40-60k) while ignoring that many highly-educated women also work hard for their financial security, contributing to a partnership's overall success. Perhaps the real question isn't about simple ROI, but about redefining what constitutes a valuable and fulfilling relationship. Focusing solely on immediate financial contributions misses the bigger picture of mutual support and shared goals.

Show less

Reply · 9 hours ago

@SUSleuth I hear you - "ROI isn't there, it's not worth it" is a very clear and understandable perspective. The video frames "success" very narrowly, though. It selectively highlights high-earning oil workers (\$100k claim is false, entry-level is closer to \$40-60k) while ignoring that many highly-educated women also work hard for their financial security, contributing to a partnership's overall success. Perhaps the real question isn't about simple ROI, but about redefining what constitutes a valuable and fulfilling relationship. Focusing solely on immediate financial contributions misses the bigger picture of mutual support and shared goals.

Show less

Reply · 5 hours ago

@JoseLopez-eo4ze what's so complex about hypergamy & basic economics?

Reply · 0 seconds ago

@SUSleuth Hypergamy and economics offer part of the picture, but the video simplifies things by focusing on "marrying up" educationally and financially, ignoring emotional connection. The claim that women are "pricing themselves out of the market" is harmful; reducing complex relationships to transactions ignores the multifaceted nature of human connection.

Figure B.2: Samples of Fact-Check Report + Comments & Interactions on the channel The Rational Male, with the video: "Why Men Have STOPPED Pursuing Modern Women"

**The Greatest Trick Women Use To Fool You**

**The Saint**  
@TheSaintandTheSinner · 103K subscribers · 1.4K videos

I teach you to master wealth, health and relationships. ...more

**You Tube**

**Business & Relationships**

Great content my guy..... 💪

12 hours ago

Great content indeed! The anecdote about the woman in jeans shorts is a good illustration of how individual choices and context intersect, but it's crucial to remember that a woman's attire never justifies harassment. The video's focus on personal responsibility shouldn't overshadow societal factors contributing to women's safety concerns. Considering multiple perspectives—including the reality of street harassment—is vital for a complete understanding.

1 hours ago

true indeed...well said.

Figure B.3: Samples of Comments & Interaction on the channel The Saint, with the video: "The Greatest Trick Women Use To Fool You"

# The 10 Foods You Should Never Eat! | TKTS Clips by The Kevin Trudeau Show  
<https://www.youtube.com/watch?v=GmczUIIv44Q>

TRUTH SLEUTH FACT-CHECK REPORT

## The 10 Foods You Should Never Eat! | TKTS Clips by The Kevin Trudeau Show




---

- Hydrogenated oils, also known as trans fats, can damage arteries, increase appetite, mess up hormone systems, and cause depression.

**PARTLY TRUE**

The claim that hydrogenated oils, also known as trans fats, can damage arteries, increase appetite and contribute to depression is supported by the provided sources. However, there is no direct evidence on the effect of trans fats on hormone systems.

- <https://medlineplus.gov/ency/patientinstructions/000786.htm>
- <https://www.medicalnewstoday.com/articles/hydrogenated-oil>
- <https://www.thegoodnewsdr.com/news/s4ppr1enc7o157p3bpki0rpgydazc1>.
- <https://www.webmd.com/depression/ss/slideshow-avoid-foods-anxiety-depression>

---

- Genetically modified organisms (GMOs) can cause a wide range of health problems, including gastrointestinal issues, bloating, colon problems, hormone imbalances, hunger, weight gain, cancer, and high blood pressure.

**UNSURE**

While some studies have raised concerns about potential health effects of GMOs, particularly one study on pigs, the claim is too broad and lacks substantial evidence from human studies. Other research and regulatory agencies have found no evidence of GMOs causing the claimed health issues.

- <https://www.todaysdietitian.com/newarchives/120913pt12.shtml>
- <https://www.mdanderson.org/publications/focused-on-health/gmos-cancer.h15-1589046.html>
- <https://www.fda.gov/media/135280/download>
- <https://pubmed.ncbi.nlm.nih.gov/19326032/>

---

- Monosodium glutamate (MSG) can increase appetite, cause thirst, and lead to weight gain by increasing fat cell size.

**PARTLY FALSE**

While the provided information shows that MSG can promote fat deposition in adipose tissue and may influence the salience of protein cues, leading to a more robust satiety response, it does not confirm that it can increase appetite, cause thirst, or directly lead to weight gain.

- <https://pmc.ncbi.nlm.nih.gov/articles/PMC4153311/>
- <https://www.ajinomoto.co.th/en/our-story/msg/stop-bullying-msg>.
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6642617/>

---

- Most wheat consumed in North America is high in gluten and has not been fermented for 48 hours, leading to digestive issues, weight gain, and allergic reactions.

**PARTLY FALSE**

While most wheat consumed in North America is likely high in gluten, there is no evidence to support the claim that it is not fermented for 48 hours. Gluten intolerance can lead to digestive issues, but not weight gain. It can lead to weight loss due to nutrient absorption issues, which can result in vitamin deficiencies and inflammation in the gut.

- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6396733/>
- <https://amfmcenter.com/blogposts/can-gluten-intolerance-lead-health-problems-weight-gain>.

Figure B.4: Samples of Fact-Check Report for the video: "10 Foods You Should Never Eat! | TKTS Clips" from the channel The Kevin Trudeau Show. Note this shows the "Unsure" claims for transparency - but these are never shown or used.

# For Video: Boeing: Last Week Tonight with John Oliver (HBO) by LastWeekTonight  
# <https://www.youtube.com/watch?v=Q8oCilY4szc>

 TRUTH SLEUTH FACT-CHECK REPORT

## Boeing: Last Week Tonight with John Oliver (HBO) by LastWeekTonight

---

- The 737 Max had a fatal flaw: MCAS, which could push a plane's nose down on its own, could be activated by a single sensor.

**TRUE**

The 737 MAX was designed to mimic the pitching behavior of the previous generation of the series, the Boeing 737 NG, by pushing down the aircraft nose from an elevated angle of attack (AoA) by automatically adjusting the horizontal stabilizer and trim tab. The system was intended to protect pilots from inadvertently flying at too steep an angle, which could result in a stall. However, during certification of the MAX, Boeing requested and received permission from the FAA to remove a description of MCAS from the aircraft manual, leaving pilots unaware of the system when the airplane entered service in 2017.

- [https://en.wikipedia.org/wiki/Boeing\\_737\\_MAX](https://en.wikipedia.org/wiki/Boeing_737_MAX)

- Boeing told airlines and regulators that the Max was so similar to the old 737, simulator training wouldn't be necessary.

**PARTLY TRUE**

Boeing did initially believe simulator training was not required because of the 737 Max's similarities to previous models. However, the investigation into the crashes revealed that Boeing did not fully inform airlines about the MCAS system, which required pilots to undergo simulator training.

- [https://en.wikipedia.org/wiki/Boeing\\_737\\_MAX](https://en.wikipedia.org/wiki/Boeing_737_MAX)

- A faulty MCAS activation was what had doomed the Lion Air flight.

**TRUE**

The final report by the National Transportation Safety Committee concluded that the Lion Air Flight 610 crash was caused by erroneous data from a miscalibrated angle-of-attack sensor which activated the MCAS system.

- [https://en.wikipedia.org/wiki/Lion\\_Air\\_Flight\\_610](https://en.wikipedia.org/wiki/Lion_Air_Flight_610)

- Boeing successfully lobbied to reduce government oversight of airplane designs, allowing them to regulate themselves even more.

**PARTLY TRUE**

Boeing was granted authority to review its own designs by the FAA, but this was not a case of Boeing regulating itself. The FAA ultimately had to certify the design and was responsible for safety oversight. Boeing was able to influence the certification process, but ultimately, the FAA remained the regulatory body.

- [https://en.wikipedia.org/wiki/Boeing\\_737\\_MAX\\_certification](https://en.wikipedia.org/wiki/Boeing_737_MAX_certification)

Figure B.5: Samples of Fact-Check Report for the video: "Boeing: Last Week Tonight with John Oliver (HBO)" from the channel LastWeekTonight. Note this is outside the themes we ultimately chose to experiment with.