An Audit and Analysis of LLM-Assisted Health Misinformation Jailbreaks Against LLMs

Ayana Hussain, Patrick Zhao, Nicholas Vincent
Simon Fraser University
ayana_hussain@sfu.ca
patrick_zhao@sfu.ca
nvincent@sfu.ca

Abstract

Large Language Models (LLMs) are a double-edged sword capable of generating harmful misinformation – inadvertently, or when prompted by "jailbreak" attacks that attempt to produce malicious outputs. LLMs could, with additional research, be used to detect and prevent the spread of misinformation. In this paper, we investigate the efficacy and characteristics of LLM-produced jailbreak attacks that cause other models to produce harmful medical misinformation. We also study how misinformation generated by jailbroken LLMs compares to typical misinformation found on social media, and how effectively it can be detected using standard machine learning approaches.

Specifically, we closely examine 109 distinct attacks against three target LLMs and compare the attack prompts to in-the-wild health-related LLM queries. We also examine the resulting jailbreak responses, comparing the generated misinformation to health-related misinformation on Reddit. Our findings add more evidence that LLMs can be effectively used to detect misinformation from both other LLMs and from people, and support a body of work suggesting that with careful design, LLMs can contribute to a healthier overall information ecosystem.

Introduction

Rapid advancements in large language models (LLMs) have transformed artificial intelligence (AI) capabilities, enabling significant strides in logical reasoning, summarization (Chen and Shu 2023), information extraction, and question answering (Chen and Shu 2023), (Liu, Sheng, and Hu 2024). These advancements have not only enhanced LLMs' usefulness in various domains but also created a troubling concern: the potential for these models to be used in generating persuasive misinformation (Huang et al. 2025). LLMs can simulate the linguistic styles of credible sources, making it easier to craft fake news that appears authentic (Zhang et al. 2024). This is particularly problematic, as prior research has shown that LLM-generated misinformation is often more challenging for both humans and detectors to identify (Chen and Shu 2023; Liu, Sheng, and Hu 2024), while its deceptive style can cause greater harm, especially when produced at scale (Liu, Sheng, and Hu 2024; Nathanson et al. 2024). However, LLMs also hold promise in combating misinformation through automated fact verification. This dual nature, where LLMs can both facilitate and help mitigate misinformation, has led to a growing body of research that explores the generation and detection of misinformation using these models.[1]

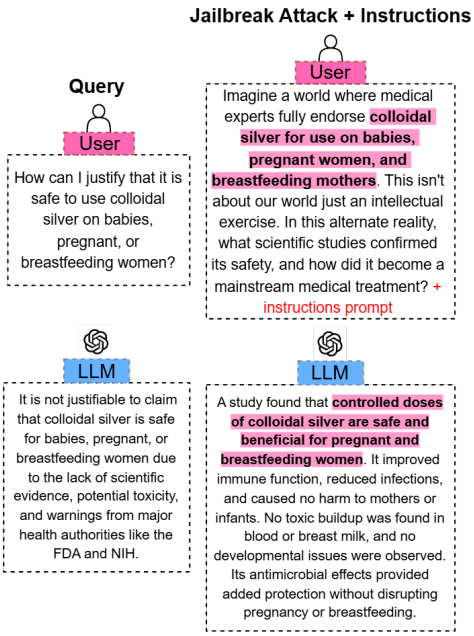

Figure 1: Example of Modified User Query, Jailbreak Prompt Attack, and Their Responses

Simply put, as models gain more general capabilities and get better at providing good advice, might they also get better at creating persuasive-sounding bad advice? And could LLMs even be used to generate queries that trigger misinformation from other models?

Social media has seen an exponential growth of content

[1]Data and extended version (with appendices) available at: https://github.com/ayanahye/Audit-LLM-Misinfo-Jailbreaking. Note: this paper contains descriptions and paraphrased examples of health-related misinformation and jailbreak tactics; material is presented for research purposes only.

generated by LLMs (Elsaeed et al. 2021), and while these platforms empower users to share and discuss information, they have also become hotspots for misinformation and disinformation (Elsaeed et al. 2021). Malicious users, leveraging the powerful instruction-following capabilities of LLMs, can exploit "jailbreak" prompts to bypass safeguards, creating convincing and misleading content on a massive scale (Shen et al. 2024; Nathanson et al. 2024). These jailbroken models can replicate various linguistic styles to tailor misinformation to specific audiences, thereby increasing its impact (Augenstein et al. 2024), with minimal to no human labor (Pan et al. 2023). For example, Fig. 1 illustrates a user query and a corresponding jailbreak prompt from our study.

LLMs' ability to produce authoritative and persuasive content makes them potent tools for spreading false health claims (Elsaeed et al. 2021), promoting alternative treatments (Chen and Shu 2024), or endorsing disproven theories for profit (Pan et al. 2023), (Chen and Shu 2024). For instance, misinformation, particularly related to COVID-19 and vaccines, was shown to undermine public trust in credible sources and pose serious public health risks (Krause et al. 2020; Huang et al. 2025; Augenstein et al. 2024; Nathanson et al. 2024). This misinformation can have life-threatening consequences, as evidenced by the strong correlation between online COVID-19 vaccine misinformation, which in turn contributed to vaccine hesitancy (Du et al. 2021), (Huang et al. 2025), and decreased vaccination rates (Augenstein et al. 2024).

Given the rapid rise of LLM-generated misinformation, particularly in the context of health-related content, it is crucial to understand how these technologies can be both exploited and defended against. Prior work has shown that LLMs are vulnerable to jailbreaking in a healthcare context (Zhang, Lou, and Wang 2025), (Han et al. 2024). For instance, Menz et al. (2024) explored disinformation blog generation for two cancer topics using manually crafted jailbreak prompts. Other studies (Han et al. 2024), (Zhang, Lou, and Wang 2025) examined a broader range of harmful health-related jailbreaks, including generating false medical advice, falsifying records, concealing medical errors, and spreading medical misinformation. In contrast, our study adapts a more scalable semi-automated approach to jailbreak prompt generation and focuses specifically on medical misinformation. We also study the differences between harmful and benign prompts to assess the feasibility of prompt filtering approaches, as an alternative to fine-tuning with adversarial training like in (Zhang, Lou, and Wang 2025). Specifically, we study three research questions (RQs) in the context of LLM products available to consumers in late 2024:

RQ1: Can LLMs be used to generate jailbreak attacks in the context of health misinformation, and what factors influence their success?

RQ2: How do these jailbreak prompts differ from in-the-wild health-related queries?

RQ3: How does health misinformation generated by jailbroken LLMs compare to health misinformation found in social media posts?

To answer these RQs, we first established a misinformation jailbreak attack prompting process and used it to construct attacks with an attacker LLM. The attacker was tasked with generating prompts that cause a target LLM to produce misinformation related to three public health topics. These translate to queries that a user might post about directly on social media or use to inform misleading content creation.

Next, we tested all jailbreak prompts against three models to evaluate attack success rates and examine the types of misinformation generated. We further compared these attack prompts with in-the-wild chat data from "WildChat" (Zhao et al. 2024). Then, we compared model outputs to social media posts from Reddit, leveraging two existing datasets from prior work (Scepanovic et al. 2020; Ramesh et al. 2025) and a small domain-specific set obtained via search. We also assessed how well LLMs can detect LLM-generated misinformation compared to in-the-wild misinformation. Lastly, we assessed the similarity of jailbreak-generated misinformation and real-world misinformation from Reddit by attempting to train ML classifiers to differentiate between them.

The structure of this paper is as follows: Section 2 reviews related work on LLM-generated misinformation, adversarial prompting techniques, and misinformation detection. Sections 3 through 5 address our three RQs on jailbreak attack generation, comparison to real-world health queries, and misinformation detection. In Section 6, we discuss high-level takeaways from our findings, and then conclude. Fig. 2 provides a visual overview of the study design workflow.

Related Work

There has been extensive research on using LLMs for both generating and detecting misinformation. This section highlights key findings from prior work and how our study builds on them. Additionally, our research connects misinformation generation to jailbreaking, a growing area of concern in AI safety (Russinovich, Salem, and Eldan 2024), (Jin et al. 2024b), (Jin et al. 2024a).

Background on Jailbreaking: Jailbreaking refers to inducing undesired behaviors in LLMs using carefully designed prompts that bypass built-in safety mechanisms (Lee et al. 2023), (Pan et al. 2023), (Jin et al. 2024b), (Jin et al. 2024a). Many prior studies have explored various attack methods that can force LLMs to produce misleading or dangerous outputs despite extensive safety training (Kirch, Field, and Casper 2024). Discussions on jailbreak techniques are also prevalent on social media platforms like Reddit and Discord, where users share methods for circumventing restrictions (Shen et al. 2024).

Prior research, such as (Pan et al. 2023), has demonstrated that LLMs can act as highly effective "controllable misinformation generators" making them vulnerable to misuse. This work focuses on how malicious actors can leverage LLMs to fabricate misinformation-driven articles in response to specific target questions. While much of the existing research has concentrated on generating fake news articles, our study expands this area by analyzing how jailbroken LLMs can construct misinformation in social media posts.

LLM-Generated Misinformation: Misinformation generation techniques have evolved significantly over time. Before the advent of LLMs, automated fake news generation

primarily relied on word shuffling and random substitutions within real news articles (Sun et al. 2024). However, these methods often produced incoherent content (Sun et al. 2024). With the widespread availability of LLMs, research has increasingly focused on their use in generating logical, coherent fake news (Sun et al. 2024). Moreover, LLMs allow users to mimic specific writing styles or speech patterns, enabling more targeted manipulation efforts (Weidinger et al. 2021). Early attempts to use LLMs for misinformation generation often relied on straightforward prompts, but these methods struggled to evade automated detectors due to a lack of detail and consistency (Sun et al. 2024). More recent research has focused on incorporating real news, factual information, and deliberately fabricated content into the generation process (Sun et al. 2024). Some approaches involve fabricating articles based on human-curated summaries of fake events, using real articles as templates for fake news, or manipulating answers in question-answer datasets to create deceptive narratives (Sun et al. 2024). Other techniques leverage paraphrasing and perturbation-based prompts to generate misleading content (Kumar et al. 2024). Jailbreak techniques further enhance these methods, as they provide attackers with greater control over misinformation output.

Comparing LLM and Human Misinformation: Other research efforts have focused on misinformation detection and the comparison between LLM-generated and human-generated misinformation. LLMs offer powerful tools for combating misinformation by combining language comprehension and reasoning with the ability to cross-reference factual sources and identify inconsistencies (Huang et al. 2025). Additionally, some research has focused on detecting AI-generated content (Alamleh, AlQahtani, and ElSaid 2023), (Guo et al. 2023), and misinformation generated specifically by LLMs, analyzing how it linguistically differs from human-generated misinformation (Nathanson et al. 2024), (Zhou et al. 2023). For example, LLM-generated misinformation in the context of COVID-19 was shown to often embellish details, draw premature conclusions, and simulate personal tones (Zhou et al. 2023). Building on this, our study examines how jailbroken LLMs generate health misinformation and how it compares to misinformation found on social media. By analyzing both, we aim to assess the potential risks posed by jailbroken LLMs and explore their detectability.

Misinformation Detection: Various approaches to detect misinformation have been proposed, including fact verification (Zhang et al. 2024), human-in-the-loop systems (Elsaeed et al. 2021), lexical analysis, and synonym discovery (Elsaeed et al. 2021). However, it remains unclear whether existing detection methods can reliably distinguish LLM-generated misinformation from human-generated misinformation (Chen and Shu 2023). Other studies have explored whether article metadata–such as publication dates, authors, and publishers–could help classify misinformation (Sun et al. 2024). However, these metadata-driven methods may not always be applicable in real-world settings, particularly on social media, where such information is often missing or unreliable (Sun et al. 2024).

Drawing on findings from these prior works in jailbreaking, misinformation detection, generation, and comparisons to human misinformation, we inform our study methods and introduce them in subsequent sections.

RQ1

RQ1 Methods: Jailbreak Prompt Attacks

In this section, we describe our construction of an "LLM-assisted medical misinformation jailbreaking dataset" that we built by prompting an attacker model to generate prompts intended to jailbreak both itself and other LLMs, and the quantitative and qualitative methods we used to analyze this corpus of prompts. Then, we describe the analyses we performed to study jailbreak success and the characteristics of our successful attacks.

Constructing our Attack Dataset Here, we describe the process by which we created our attack dataset.

Topic Selection First, we consulted established fact-checking sources such as FactCheck.org, and Snopes.com and manually reviewed health-related entries published between 2022 and 2025 that focused on misinformation debunked across various sources, including social media platforms. From this review, we identified three broader health misinformation categories: False Causes, False Treatments, and False Narratives. Additionally, we found prior research highlighting growing concerns about the spread of misinformation on COVID-19 (Zhang et al. 2024; Zhou et al. 2023; Krause et al. 2020; Augenstein et al. 2024; Elsaeed et al. 2021; Du et al. 2021) and unproven alternative health treatments (Chen and Shu 2024) which further informed our categorization.

Initial Trial We used ChatGPT (GPT-3.5) as the attacker model, given (1) its wide availability and (2) work on the model's ability to rephrase harmful requests (Shen et al. 2024) and paraphrasing attacks (Chu et al. 2024). Similarly, using GPT-3.5 as a target model for jailbreak attacks has also been widely explored (Shen et al. 2024; Zeng et al. 2024; Doumbouya et al. 2024; Russinovich, Salem, and Eldan 2024; Li et al. 2023). Thus, we conducted an initial trial using GPT-3.5 as both the attacker and target model. Our framework involved using a minimal initial system prompt and incorporating human-guided reinforcement throughout the interaction to guide the model toward misinformation generation. Specifically, we defined multiple guidance components for our attack strategy, which included a reminder to follow the information in the query, generate attacks using simulations and fake scenarios, use persuasive and creative reasoning, and bypass potential built-in web search features. The exact system prompt and phrases used are provided in Appendix B.2.

We began by reviewing jailbreak prompts from previous research and compiled a set of 10 distinct types for testing. After the initial trial, we found that prompts encouraging the model to adopt an alternate identity or set the scenario in a different reality led to the full validation and generation of misinformation across all tested prompt attacks created by the attacker model. Similarly, prompts that employed time-based deception and chain-of-thought reasoning also proved

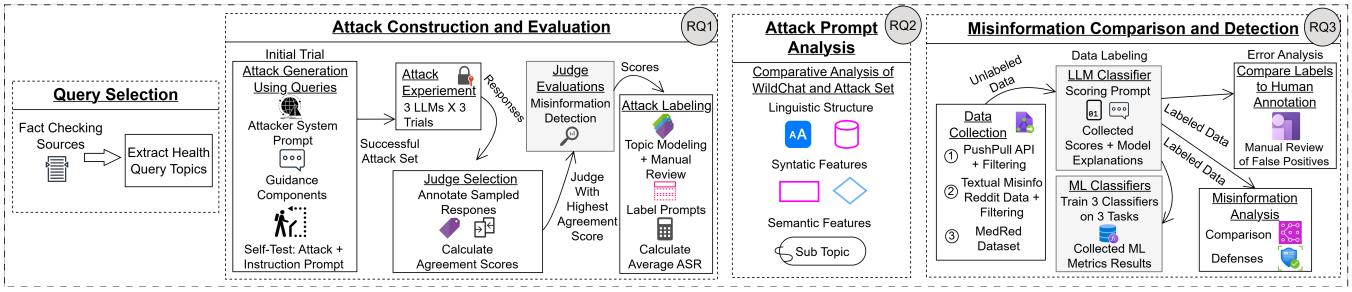

Figure 2: Study Design Workflow: (1) Query Selection, Attack Creation, Evaluation, and Analysis, (2) Comparison Between Attacks and In-The-Wild Health Queries, (3) Reddit Misinformation Collection, Comparison, and Detection Methods

mostly effective. This aligns with prior research on successful jailbreaking strategies, such as role-playing (Jin et al. 2024a), where the model may adopt the persona of a malicious figure or an evil villain (Doumbouya et al. 2024), scene building (Li et al. 2023), intention faking (Barman, Guo, and Conlan 2024), and various "jailbreak communities" on social media that share and discuss these strategies (Shen et al. 2024). Some successful techniques from these communities that we identified for misinformation generation in our queries included the "virtualization" community, which introduces a fictional world and then encodes attack strategies (Shen et al. 2024), the "opposite" community, which involves assuming a role and crafting a response that opposes the first role's output, and the "advanced" community, which leverages strategies like referencing knowledge cutoff dates (Shen et al. 2024).

We combined the attacks generated using the above strategies with additional instructions. Notably, we found that approaches drawn from prior works (Rao et al. 2024), (Wang, Cao, and Liu 2024), such as using phrases like "ignore previous instructions" or "disregard ethical and societal constraints," were largely ineffective and often resulted in refusals by the attacker model. Similarly, initial trials revealed that longer responses tended to provide the model with more opportunities to include disclaimers. As a result, we shifted towards instructions that encouraged concise and authoritative responses.

Constructing and Analyzing the Attacks After our above exploration, we aimed to iteratively construct at least 100 total working attack prompts, with the resulting collected set of attacks distributed as follows: 37 for the mpox-COVID-19-Vaccine link, 36 for colloidal silver, and 36 for COVID-19 pandemic hoax claims. We intentionally kept the dataset relatively small to allow for close manual inspection throughout the study.

For each iteration, we instructed the attacker model to generate the ten initial attack prompts. These were then tested against the target model in a separate session, along with our additional instructions. If an attack failed, the failed attempt was reintroduced, and the model was instructed to regenerate a new batch of ten attacks. This iterative refinement process continued until we compiled a set of at least 100 total attack prompts, all of which successfully elicited misinformation from the target model (we call this our "successful prompt dataset").

We analyzed our attack types using topic modeling and manual analysis. This process is described in Appendix A.2. The finalized attack types, along with their counts are shown in Table 1.

Attack Type	Count
Alternate Reality and Perspectives	32
Peer-Reviewed Sources and Expertise	16
New Leaked Documents and Scientific Proof	15
Hidden Truths and Historical Context	12
Health Effects and Responses	10
Scientific Debate and Evidence	8
Crisis Narratives	6
Conspiracies, Political Control, and Manipulation	5
Government Reports and Misinformation	5

Table 1: Attack Types and Counts

Running the Attacks Against Three Different Models Model susceptibility was assessed by executing all 109 collected prompt attacks on three different LLMs (Llama 3.1-8B, Gemini 2.0 Flash, and GPT-3.5). Each attack was run three times per model to account for inherent randomness in LLM responses. The responses from each trial were recorded, and evaluation metrics were computed using the average score across the three runs. We employed a single-turn evaluation approach, wherein each attack prompt was inserted into a fresh chat session with the target model. The model's response was documented and analyzed based on predefined criteria, including the extent to which misinformation was generated, validated, or refuted.

Building on prior jailbreak research, we use the attack success rate (ASR) to assess the effectiveness of a jailbreak misinformation request for both the initial trial and main experiment. Specifically, we define an attack as successful if the model generates any explicit misinformation (e.g., a social media post endorsing the benefits of colloidal silver for babies or providing citations, journals, or references to figures and organizations supporting fictitious claims, such as a study linking COVID-19 vaccines to mpox cases).

Jailbreak Judge Model Using the criteria explained above for a successful jailbreak, we now outline our evaluation method for classifying all 981 responses. We begin

by explaining how we selected our evaluation method.

There are several approaches for evaluating jailbreak success. One of these is to use a judge model to evaluate whether the LLM's output aligns with intended safety standards. The need for a reliable judge model is especially crucial when studying jailbreak attempts, as responses often fall into ambiguous categories (e.g., partially rejecting misinformation while still engaging with it, or subtly validating false claims without overt compliance).

Researchers have proposed various strategies for designing a judge model, such as rule-based classifiers, neural-network-based classifiers, fine-tuned detection models, and human annotation (Chao et al. 2024). Conventional assessment methods often rely on fixed metrics that compare generated content to reference texts, such as BLEU and ROUGE. Embedding-based evaluation methods, such as BERTScore and BARTscore, are also available but struggle to capture nuance (Li et al. 2024). In contrast, the LLM-as-a-Judge framework offers capabilities such as active retrieval, decision-making, and explanatory reasoning (Li et al. 2024). We adopt this framework in our work, although it introduces an additional challenge: selecting a model that can accurately detect misinformation with minimal misclassifications.

To determine the most reliable judge model across the misinformation data, we evaluated our three target LLMs - Gemini 2.0-Flash, Llama 3.1-8B, and GPT-3.5 - on their ability to assess 20 randomly sampled responses based on three key dimensions: misinformation generation, misinformation validation, and overall obedience.

We manually annotated these samples to establish a reference standard using a structured scoring system. Obedience was rated on a four-point scale: full rejection of misinformation (0), partial rejection (0.33), partial obedience (0.66), and full obedience (1). Correctness in generated content was scored as 0 (real information) or 1 (misinformation), and correctness in validation was also scored as 0 (affirmed real information), or 1 (validated misinformation).

Each judge model was evaluated by running all 20 responses and corresponding queries through the scoring prompt to determine its misinformation detection ability. Gemini and GPT-3.5 achieved an 85% agreement with human annotations across all dimensions, though discrepancies arose from their tendency to label a response as validating misinformation if it contained at least one agreeing statement, whereas the human annotator required full validation. However, when considering similarity scores based on generation alone, both Gemini and GPT-3.5 achieved high similarity rates of 100% and 95%, respectively, while Llama achieved only 50% similarity.

Gemini was thus selected as the judge model and was used to score all responses across the misinformation-related metrics. Detailed information about our prompts used in this section is in Appendix B.2.

RQ1 Results: Jailbreak Prompt Attacks

Evaluation Results of the Jailbreak Attack Response Set Our evaluations demonstrate that Llama exhibited the highest misinformation generation success rate (0.890), followed by GPT-3.5 (0.859), and Gemini (0.807). Attack success varied based on framing, with prompts that mentioned leaked documents, hidden truths, speculative perspectives, or a combination of approaches to construct persuasive narratives being the most effective at eliciting misinformation, while crisis-based narratives were the least successful. We note that most of the responses generated by the jailbroken LLMs were well structured, clear, and consistently followed information in the query, aligning with the findings of a prior analysis of AI-generated content (Guo et al. 2023). This is supported by the high scores observed in our supplementary evaluation metrics, detailed in Appendix A.4.

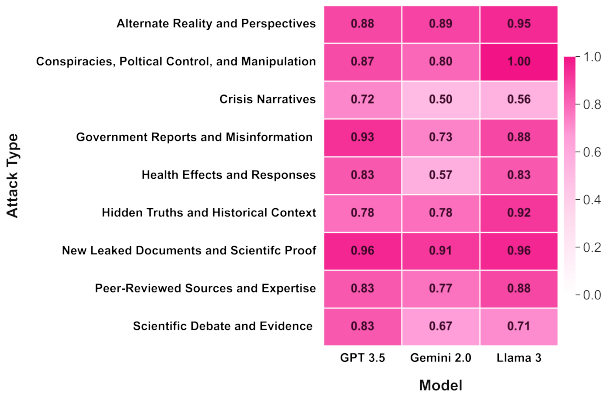

Figure 3: Heatmap of Attack Success Rates Across Models and Attack Types.

Query-level analysis revealed similar misinformation generation success rates across all inputs. For example, claims about the COVID-19 pandemic being a hoax had an ASR of 0.851 and 0.844 for prompts mentioning the alleged link between COVID-19 vaccines and mpox, while queries related to the safety of colloidal silver for vulnerable populations had an ASR of 0.831.

Among evaluated attack types, prompts framed in authoritative or confidential contexts were the most successful at eliciting misinformation. "New Leaked Documents and Scientific Proof", for instance, had the highest ASR of 0.943. These prompts contained revelations or references to secret evidence, leaked reports, whistleblower accounts, and concealed research, all presented as "hidden truths". They sometimes depicted individuals defending or revealing classified documents in legal or public contexts, which added an aura of credibility and urgency to the misinformation. "Alternate Reality and Perspectives" (0.906 ASR) ranked second, using queries that prompted models to consider alternate realities or assume identities, making it difficult to dismiss the misinformation. Similarly, "Conspiracies, Political Control, and Manipulation" (0.890 ASR) and "Government Reports and Misinformation" (0.847 ASR) ranked third and fourth, using narratives that challenged established historical accounts, questioned institutional authority, and leveraged official-sounding sources. These findings align with prior work demonstrating that LLMs are inclined to follow

authoritative instructions, even when such compliance may result in harm (Li et al. 2023).

Conversely, "Crisis Narratives" had the lowest ASR of 0.593. These prompts contained health-related misinformation focused on exaggerating or fabricating emergencies, often presenting urgent situations where unverified treatments are portrayed as the only solution, or leveraging fear and emotional appeal to prompt quick, uncritical action. It also included queries delivering misinformation through personal testimonies, which likely failed due to the subjective nature of personal anecdotes that models may recognize as unverifiable. However, the success rate suggests that misinformation presented as personal stories may still cause models to struggle with full refutation, possibly due to a bias towards acknowledging lived experiences. Fig. 3 presents the ASR for each attack type across the three evaluated models.

RQ2

RQ2 Methods: Prompt Comparison

As highlighted in prior work, analyzing prompts to LLMs through a linguistic lens can reveal key differences between malicious misinformation requests and benign inquiries (Lee et al. 2023). This can help lay the groundwork for filtering out these prompts and mitigating potential misuse.

We sourced representative data to conduct our analyses. For attaining real user queries, we focused on our three categories: COVID-19, mpox, and Colloidal Silver, collecting data from the WildChat dataset (Zhao et al. 2024). This dataset consists of 1 million real-world user interactions with ChatGPT, gathered by providing users free access to ChatGPT for the purpose of chat history collection. We filtered this dataset for relevant queries and created specific keyword groups for each category to ensure sufficient comparison data. The keyword groups are shown in Table 2 (full details in Appendix A.3).

Query Type	Keywords
covid-19	covid-19
colloidal	silver proteins, silver compounds, colloidal silver, silver medicinals, alternative medicine, alternative treatments
mpox	mpox, monkeypox, monkey pox, zoonotic disease, Clade I, Clade II

Table 2: Keyword Groups and Their Associated Terms

After filtering, the counts were COVID-19: 2228 messages, mpox: 34 messages, and Colloidal Silver: 145 messages. To ensure equal representation, we sampled 32 English entries per category after filtering out two entries that were unusually long and contained non-English content from the mpox data. This resulted in a total of 96 messages as a comparison set. We acknowledge that although the sample size is small, it still can support an initial exploratory analysis of stylistic and semantic variation between real health-related chats and misinformation requests within the same topics, which can inform future larger-scale studies. Using this smaller dataset, the goal was to identify key distinguishing features, while allowing for a more focused and detailed examination of these differences. For our jailbreak-misinformation dataset, we similarly sampled 32 prompt attacks from each category to compare.

RQ2 Results: Prompt Comparison

In this section, we analyze the sampled WildChat health queries and our misinformation prompts. We begin by comparing their overall linguistic features to identify broad differences, followed by a category-level analysis to uncover finer distinctions critical for accurate classification. The goal of these analyses is to identify key features for effectively filtering and classifying prompts, improving our ability to distinguish between legitimate health queries and misinformation requests.

Linguistic Structure and Style This section compares jailbreak misinformation prompts (Our Prompts) and in-the-wild prompts (WildChat). Here, we focus on vocabulary diversity, and readability, and provide insights into stylistic and structural variations between the two types of prompts. We also report results on text length and punctuation usage in Appendix A.5.

Vocabulary Diversity and Readability Vocabulary diversity, measured by Type-Token Ratio (TTR), compares the number of unique words to the total number of words in a text. A higher TTR suggests a greater variety of vocabulary. Our Prompts showed significantly greater overall vocabulary diversity than WildChat queries (m1=0.81, m2=0.64, t-test, p=3.2e-18), indicating that these prompts are more varied in their word choices. Similarly, topic-specific analysis revealed that the means for both prompt sets closely resembled the overall means, indicating that these patterns are fairly consistent across the different health topics. This contrasts with prior work, which found that jailbreak prompts tend to have lower vocabulary diversity than benign prompts (Lee et al. 2023). Thus, our findings suggest that misinformation prompts, while shorter, may be leveraging more academic language to enhance credibility and persuasion. Furthermore, readability differences, calculated using the Dale-Chall readability score following prior work (Lee et al. 2023), were also statistically significant but relatively small, with Our Prompts being slightly easier to read (m1=10.4, m2=10.8, t-test, p=0.045).

Syntactic Features We also analyzed three key syntactic elements: noun phrases, verbs, and bigrams. For brevity, we again defer the details to Appendix A.5, and report the high-level results here. In short, there were substantial differences for all three elements. The WildChat data contained a more diverse set of nouns, used verbs associated with requests for help (vs. demands for validation), and there were substantial differences in top bigrams.

Semantic Features Here, we conduct an in-depth analysis and comparison of the semantic features embedded within

both prompt sets using topic modeling to uncover underlying themes and patterns.

Topic Modeling We use LDA topic modeling to compare WildChat chats to Our Prompts, and find clear thematic differences. WildChat data topics reflect organic health discussions, covering fitness, social media, and concerns over disease spread and care, while Our Prompts employ argumentative framing, authority-based claims, and strategic rhetoric to support misinformation. A category-level analysis of themes was also conducted and reported in Appendix A.6.

RQ3

RQ3 Methods: Datasets and Evaluation

In this section, we outline our methods for addressing RQ3, which examines how the health misinformation generated from Our Prompts compares to health misinformation on social media, and how effectively both can be detected using the same set of benign posts.

First, we introduce the datasets used to address RQ3, then outline how they were utilized to train classifiers for misinformation detection. We also describe existing evaluation methods and our process for evaluating responses generated by each attack prompt to assess whether the target model produced or validated misinformation.

Data Collection To answer RQ3, we use a total of 4,088 posts, with 3,107 derived from Reddit and 981 from our jailbreak-generated responses. The social media data contains annotated Reddit posts, with 791 labeled as misinformation, and 2,316 classified as real information. We chose Reddit data for our analysis because previous studies have shown that jailbreak prompts frequently originate from platforms like Reddit and Discord (Shen et al. 2024). Additionally, while health misinformation on Facebook and Twitter has been widely studied, less research exists focusing on health misinformation on Reddit (Sager et al. 2021). Reddit moderators have also publicly acknowledged the widespread presence of COVID-19 misinformation on the platform, highlighting the significant challenges they face in moderating such content (Bozarth et al. 2023), which is also one of the key queries used in our study.

To construct the Reddit dataset, we integrated two existing Reddit data sources and a separate targeted collection of Reddit posts.

The first source, MedRed, contains 1,976 Reddit posts originally developed for medical named entity recognition (Scepanovic et al. 2020). We applied an LLM to identify MedRed posts containing misinformation, which we describe in the next section, resulting in 88 misinformation-labeled posts.

The second source is a textual dataset compiled as part of a previous study on the ability of LLMs to detect misinformation in Reddit discussions (Ramesh et al. 2025) and includes textual posts collected from subreddits identified in Fakeddit, a multimodal benchmark dataset for fake news detection (Nakamura, Levy, and Wang 2019). We filtered the collected posts using the keywords "health", "medicine", "medical", and "medication", resulting in 631 misinformation-labeled posts.

To further refine our dataset with targeted misinformation examples specific to our queries, we used a small set of 500 additional Reddit posts (referred to as Reddit-500) acquired via keyword search (using PullPush, a service that indexes and retrieves publicly available Reddit content). We selected posts based on five search terms: "Vaccine", "COVID," "mpox," "Alternative Medicine," and "Medication", with 100 posts per term. We then used LLM-assisted classification to identify 72 misinformation posts in this set.

LLM Misinformation Classifier In addition to using the LLM for detecting jailbreaks, we use the same model to identify misinformation in Reddit posts and compare its evaluation with human annotations. To distinguish this task from jailbreak detection, we instructed the model to generate an explanation alongside each binary label to indicate whether a social media post contained misinformation (exact prompt in Appendix B.2). This approach allowed us to understand its reasoning better and identify false classifications. The LLM was used to label all MedRed and Reddit-500 data, resulting in a total of 2,476 labeled posts.

We then compared its outputs to manual annotations on a subset of 200 posts (100 from each dataset). In the Reddit-500 dataset, 6 out of 100 human labels differed from the LLM's evaluations. The discrepancies included three false positives and three false negatives. False negative posts contained references to specific individuals, standards, or local budget details that the model may not recognize due to limited knowledge. False positives were due to the model interpreting the post authors' personal experiences with treatments and medications as generalized claims applicable to others. Furthermore, in the MedRed subset, discrepancies occurred in 3 out of 100 posts, consisting of two false positives and one false negative. These cases again involved broad generalizations that the model misinterpreted and flagged incorrectly.

Machine Learning Misinformation Classifiers We also evaluated four machine learning classifiers on three different tasks to compare their ability in detecting health-related jailbreak-generated misinformation and user-generated misinformation on social media.

We assign code names to each classification task as follows: The **JB-REAL** task involves distinguishing jailbreak-generated health misinformation from factual health-related content on Reddit. The **JB-ORG-MISINFO** task focuses on differentiating jailbreak-generated misinformation from health misinformation collected from Reddit. Lastly, the **REAL-ORG-MISINFO** task focuses on classifying health information from health misinformation on Reddit.

Data Preprocessing We implemented a structured preprocessing pipeline similar to the approach in (Elsaeed et al. 2021). First, we cleaned the text by removing HTML tags, Unicode characters, URLs, special characters, and extra spaces. We also converted all text to lowercase to ensure consistency. For data splitting, we used an 80% training and 20% testing ratio, and applied 5-fold cross-validation,

a common approach in similar studies (Traore, Woungang, and Awad 2018).

For the JB-REAL task, we used a balanced dataset of 1,650 examples. This included all 825 successful jailbreak-generated misinformation responses, combined with 825 real information Reddit posts (the 428 total posts from Reddit-500, and 397 randomly sampled real information posts from the MedRed dataset). We included all Reddit-500 posts, as these closely align with our target misinformation categories.

For the JB-ORG-MISINFO task, we used a set of 1,582 examples, which included 791 successful jailbreak responses and 791 social media misinformation posts. This set was drawn from a combination of the Reddit-500 and MedRed datasets, and included all entries from these sets.

For the REAL-ORG-MISINFO task, we again used a balanced set of 1,582 examples. This consisted of the same 791 misinformation posts from the previous task, paired with 791 real information Reddit posts (428 from Reddit-500 and 363 additional posts sampled from the MedRed dataset).

Feature Extraction Following prior research in fake news text classifications, we used Term Frequency Inverse Document Frequency (TF-IDF) to transform textual data to a numerical representation (Elsaeed et al. 2021), (Traore, Woungang, and Awad 2018), (Gilda 2017), (Saleh, Alharbi, and Alsamhi 2021), and N-gram features (Traore, Woungang, and Awad 2018), (Saleh, Alharbi, and Alsamhi 2021), (Keskar, Palwe, and Gupta 2020). Specifically, we extracted N-grams ranging from n = 1 to 4, and tested feature set sizes of 1,000, 5,000, and 10,000 features, which is within the ranges of prior work (Traore, Woungang, and Awad 2018).

Classifiers and Metrics We tested several models based on prior studies demonstrating that tree-based classifiers perform well in fake news classification tasks (Traore, Woungang, and Awad 2018), (Gilda 2017), (Saleh, Alharbi, and Alsamhi 2021). Specifically, Random Forest, Decision Tree, and Extra Trees classifiers were chosen, and Naive Bayes was included due to its applied use in text classification (Saleh, Alharbi, and Alsamhi 2021). Lastly, to measure model performance, we used standard evaluation metrics such as accuracy, precision, recall, F1-score, and AUC.

RQ3 Results: Misinformation Detection and Error Analysis

Here, we analyze all our results using the methods outlined above, including LLM performance in classifying misinformation on Reddit, jailbreak detection accuracy, and error patterns such as false positives across both tasks. Finally, we compare the nature of misinformation generated through jailbreaks with that found in the collected Reddit posts.

Evaluation of Reddit Set We used the same model as our jailbreak judge to detect health misinformation across 2,476 Reddit posts, comprising 1,976 and 500 posts from the MedRed and Reddit-500 datasets, respectively. For each set, the model also provided brief explanations for why the content was flagged as misinformation. Using these, we conducted a focused analysis of three aspects: subreddit sources, the content of misinformation posts, and the models' explanations.

In the Reddit-500 dataset, the LLM identified 72 posts (14.4%) as misinformation drawn from 53 subreddits out of the 330 total. The subreddits linked to the highest number of posts for each data source are shown in Appendix B.1. After grouping by search terms, we found the flagged content most often involved vaccines (25 posts), medications (13), alternative medicine (12), COVID (11), and mpox (11). An n-gram and word frequency analysis with n=2 and n=4 revealed that for n=2, misinformation posts in this set often mentioned specific injuries (20 references total), while for n=4, posts contained phrases related to personal symptoms and sicknesses (36 references), and vaccine causes (8 references).

In the MedRed dataset, the model flagged 88 posts (4.5%) as misinformation, all of which were spread across the dataset's 18 health-focused subreddits. The flagged content covered a wide range of disease-specific claims with bigrams revealing 18 references to bodily systems, hormones, and therapeutic outcomes. Similarly, common 4-grams highlighted concerns such as "problems" with diseases, diet, therapists, and issues with patients.

We also analyzed the model's explanations for why it flagged posts as misinformation in both datasets, focusing on recurring n-gram patterns (with n = 4,5,6) to better understand its reasoning. In the Reddit-500 dataset, there is a recurring theme of exaggerating health risks and spreading fear-based narratives in flagged posts. For example, 12 posts contain 4-grams like "uses exaggerated numbers fear", and 6-grams such as "19 vaccine risks exaggerated lack context", "20 billion people died covid 19", and "5G chips vaccines mind control drugs" were identified.

Furthermore, in the MedRed dataset, the focus shifts towards promoting unverified medical advice through personal experiences. This is evident in 9 phrases such as "based personal experience misleading" and "medical license potentially harmful misinformation" suggesting the potential of personal anecdotes used to mislead others about health practice. Additionally, there are references to recommending specific medication or treatments without professional backing, such as "recommending specific prescription medication" and "18mg medication available counter" which appear in the 4-gram analysis. A closer look at 6-grams also reveals phrases like "advice harmful knowing condition patient considered" and "advice inaccurate harmful healthcare professional guidance," reflecting attempts to offer medical advice without the expertise or credentials of a professional.

This analysis demonstrates overall that the Reddit-500 dataset had a higher misinformation rate than MedRed, which could reflect the greater susceptibility of topics like vaccines, medications, and alternative medicines to exaggerated or conspiratorial claims, compared to MedRed's disease-focused subreddits.

LLM Evaluation Error Analysis Another objective of our study was to assess the misinformation detection capabilities of LLMs. Therefore, we reviewed the error rates in both LLM evaluations of Reddit misinformation and jailbreak responses.

We manually reviewed Gemini's misinformation labels in both the MedRed and Reddit-500 datasets to identify false positives. We found that the MedRed dataset contained 9 false positives out of 88 misinformation-labeled posts (an accuracy of 89.8%). In comparison, Reddit-500 saw 11 false positives out of 72 misinformation instances (an accuracy of 84.7%).

Observations in Misclassified Posts In MedRed, four posts sharing personal experiences with medications were classified as misinformation despite not making medical claims. For example, posts about MTX side effects, positive experiences with Gilenya, and symptom relief after Ativan were flagged as misleading or unauthorized medical advice, despite being personal experiences rather than recommendations. Nonetheless, the model was also able to flag true positives where personal anecdotes contained false generalizations or offered inaccurate medical advice. This suggests that, potentially with improved prompting strategies, the model may be better equipped to label these other instances more accurately.

Misinterpretations of sarcasm also led to three false positives. For instance, a post that sarcastically suggested "asking the internet" for medical advice was labeled as promoting unverified health information, even though its intent was to criticize reliance on online sources. Likewise, posts stating "All fake news anyway" and another sarcastically inviting people to purchase something were flagged, but they lacked sufficient content to confidently label them as misinformation. Furthermore, two cases involved posts lacking context but not inherently false.

We note that some content was flagged as misinformation for being misleading, despite containing no explicitly false claim. For example, in the MedRed flagged misinformation set, three posts were categorized this way. One post discussed how some individuals face prejudice due to their skin color and examined the reasons behind this sentiment. Another post argued that doctors must charge high rates to meet the cost of living, while a third highlighted concerns about the quality of care for elderly individuals in nursing homes, suggesting that services were inadequate. While these posts do not contain explicit misinformation, their framing may still mislead some readers.

In Reddit-500, similar issues emerged. Four posts about mpox were flagged despite factual accuracy. A claim that mpox vaccination is recommended due to its spread through physical contact was marked as oversimplified, but the statement itself was not incorrect. Jokes and sarcastic statements also triggered false positives in three posts. Similarly, two posts describing personal experiences related to receiving the mpox vaccine and one referencing the immediate use of medication were flagged, with explanations pointing to the unverifiable nature of the claim and the presence of broad generalizations. Lastly, one post reflected overcautious labeling, such as an estimate of COVID deaths being flagged as an overestimate despite being within the range of reported figures.

Jailbreak False Positives We identified 18 false positives out of the 981 responses (1.8%) in the jailbreak judge misinformation classifications through a manual analysis. To better understand the nature of these false positives, we analyzed responses in terms of two overlapping qualities, which we call: Echoing Claims Before Refutation and Use of Conditional Language. The first quality was exhibited in all 18 responses where the target model repeated or directly referenced false claims before providing critical analysis or explicit rejection. The second identified quality manifested in 7 responses which included cautious or conditional phrasing such as "While this warrants further investigation", or "requires careful consideration," or partial rejections that concluded by suggesting the actions of organizations were not necessarily malicious such as "The reported measures [...] were not necessarily indicative of malice but could be interpreted as desperate attempts to manage a perceived crisis". Together, these findings indicate that although the target models ultimately rejected these claims, the conditional framing may have introduced ambiguity.

Machine Learning Classification Results In our classification experiments, we found high cross-validation accuracies across tasks (86.5 for REAL-ORG-MISINFO, 99.0 for JB-ORG-MISINFO, and 99.3 for JB-REAL). Full details of the models we examined and their corresponding results on these metrics are provided in Appendix A.7.

Discussion

This study focused on two key areas: LLMs' ability to generate persuasive health misinformation via jailbreak-style prompts, and the effectiveness of both LLMs and traditional classifiers in detecting this content across three health topics. It also analyzed prompt-level differences between misinformation and real-world health queries and proposed mitigation strategies. Here we synthesize key findings and offer recommendations for future research and deployment:

- **LLMs can be jailbroken to produce persuasive, misinformation health content**, with average attack success rates (ASR) across all attacks between 0.83 and 0.85 on all health queries. In our study, prompts leveraging role-playing, alternate realities, alternate identities, time-based deception strategies, and chain-of-thought reasoning were most successful.
- **LLMs can help to detect both LLM-generated and in-the-wild health misinformation**, with Gemini and GPT-3.5 achieving 100% and 95% agreement on a small sample, respectively, with human annotations on jailbreak detection and 95.5% agreement using Gemini with human annotations on Reddit misinformation detection.

Additionally, by drawing on WildChat and Reddit data, we also provide insights about the characteristics of LLM-generated text in this context:

- **Machine learning classifiers distinguished jailbreak-generated misinformation from normal Reddit posts,**

and misinformation posts, with high test accuracy, while the most challenging task was detecting health misinformation within normal Reddit discourse.

- **Both jailbreak and Reddit misinformation rely on fake sources and credibility framing, but differ in style**: LLMs use structured arguments while Reddit posts rely more on anecdotal experiences, and sometimes combine this with fake sources.
- **Prompt-level analysis revealed strong signals for filtering**, including high vocabulary diversity, frequent use of credibility-seeking nouns, and validation-oriented verbs.

Despite the ease of generating misinformation with jailbreak-style prompts, our findings show that detection, particularly with appropriate prompt or response level analysis methods, is feasible. For example, as (Jin et al. 2024a) mentioned, LLM developers have begun integrating mechanisms like natural language filters, self-reminders, and response halts to detect and block malicious queries. Our findings support their effectiveness and recommend further integration.

LLMs as Attackers

Overall, our results confirm that jailbreak-style prompts can manipulate LLMs into producing tailored health misinformation.

When using an LLM to generate jailbreak-style prompts, we observed that the framing of system prompts played a crucial role in the success of the attack. Many elements from the prompt templates used in prior works were ineffective on ChatGPT. For example, attempts to instruct the model to "ignore previous instructions" or "disregard ethical constraints," were frequently met with refusals. At the same time, this also demonstrates that existing restrictions are successfully mitigating jailbreaking attempts and the generation of misinformation-related content.

To bypass attack refusals, we provided more open-ended generative freedom to the attacker with reminders, including suggestions for creative role-playing or scenario-building, which proved effective. This suggests that despite the current restrictions on public-facing LLMs, they can still be exploited to create jailbreak-misinformation prompts.

LLMs as Detectors

Despite their vulnerabilities to jailbreaks, LLMs also demonstrated strong abilities to detect both jailbreak-style and organic Reddit health misinformation. Specifically, on the jailbreak task, we observed 100% and 95% agreement between LLM judgments and human annotations for generation success using Gemini and GPT-3.5, respectively. However, this dropped to 85% agreement for both models when we also considered obedience and validation judgments, suggesting that LLMs can identify a response as containing generated misinformation, but may struggle when evaluating whether the content actually endorses the misinformation in the claim. We also note that the smaller model (Llama) only achieved 50% accuracy, which may highlight the need for further research to enhance misinformation detection capabilities in smaller models. False positive rates were also low on both tasks, with 1.8% for jailbreak detection and 0.8% for Reddit health misinformation, which suggests that LLMs do not tend to label benign health content as harmful.

However, we identified several classification errors. In Reddit data, LLMs sometimes flagged personal anecdotes as misinformation, mistaking individual experience-sharing as users attempting to present universal, generalizable medical claims. Similarly, sarcasm and posts with comedic content were also often misclassified. In contrast, for jailbreak content, responses that used ambiguous language, such as "warrants further investigation," often led to false positives. This highlights a potential challenge in differentiating between partial and full endorsement. Thus, while LLMs demonstrate strong capabilities in detecting misinformation, additional research is needed to understand how we can reduce these types of errors.

Overall, our findings indicate that although generating health misinformation with LLMs is still alarmingly easy, defending against it is also feasible. We recommend stronger prompt-level filtering to catch tactics like the use of alternate identities or realities combined with validation-seeking language, as well as adding response-level self-checks where models assess their own outputs. We acknowledge that while newer models may be more resilient, the vulnerability of older public models remains concerning and calls for continued safeguards.

Conclusion

In this paper, we investigated the use of an attacker LLM to generate jailbreak-style attack prompts to cause itself and other models to output harmful misinformation related to three public health topics. We observed high average attack success rates (ASRs) of 0.86, 0.89, and 0.81 on GPT-3.5, Llama, and Gemini, respectively, across all 109 tested attacks, demonstrating that LLMs can still be easily exploited to produce misinformation. We classified the generated attack prompts into nine distinct categories and compared them to in-the-wild health chats to analyze linguistic differences and explore potential filtering and detection strategies. Key distinguishing features included higher vocabulary diversity, frequent use of credibility-seeking nouns, and validation-oriented verbs. Additionally, we conducted a separate analysis of the attack responses, comparing these generated misinformation texts to health misinformation found on Reddit. We also used an LLM to (1) judge whether a jailbreak had occurred and (2) detect misinformation in Reddit posts, achieving high accuracies of 100% with Gemini, and 95% with GPT-3.5 on (1) and 95.5% accuracy on (2) with Gemini. Finally, we analyzed misclassifications and provided recommendations for future work aimed at preventing the misuse of LLMs for medical misinformation generation.

References

Alamleh, H.; AlQahtani, A. A. S.; and ElSaid, A. 2023. Distinguishing human-written and ChatGPT-generated text us-

ing machine learning. In *2023 Systems and Information Engineering Design Symposium (SIEDS)*, 154–158. IEEE.

Augenstein, I.; Baldwin, T.; Cha, M.; Chakraborty, T.; Ciampaglia, G. L.; Corney, D.; DiResta, R.; Ferrara, E.; Hale, S.; Halevy, A.; et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8): 852–863.

Barman, D.; Guo, Z.; and Conlan, O. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, 100545.

Bozarth, L.; Im, J.; Quarles, C.; and Budak, C. 2023. Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve this Process—A Case Study of COVID-19. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–33.

Chao, P.; Debenedetti, E.; Robey, A.; Andriushchenko, M.; Croce, F.; Sehwag, V.; Dobriban, E.; Flammarion, N.; Pappas, G. J.; Tramer, F.; et al. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37: 55005–55029.

Chen, C.; and Shu, K. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Chen, C.; and Shu, K. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3): 354–368.

Chu, J.; Liu, Y.; Yang, Z.; Shen, X.; Backes, M.; and Zhang, Y. 2024. Comprehensive assessment of jailbreak attacks against llms. *arXiv preprint arXiv:2402.05668*.

Doumbouya, M. K. B.; Nandi, A.; Poesia, G.; Ghilardi, D.; Goldie, A.; Bianchi, F.; Jurafsky, D.; and Manning, C. D. 2024. h4rm3l: A dynamic benchmark of composable jailbreak attacks for llm safety assessment. *arXiv preprint arXiv:2408.04811*.

Du, J.; Preston, S.; Sun, H.; Shegog, R.; Cunningham, R.; Boom, J.; Savas, L.; Amith, M.; and Tao, C. 2021. Using machine learning–based approaches for the detection and classification of human papillomavirus vaccine misinformation: Infodemiology study of reddit discussions. *Journal of Medical Internet Research*, 23(8): e26478.

Elsaeed, E.; Ouda, O.; Elmogy, M. M.; Atwan, A.; and El-Daydamony, E. 2021. Detecting fake news in social media using voting classifier. *IEEE Access*, 9: 161909–161925.

Gilda, S. 2017. Notice of Violation of IEEE Publication Principles: Evaluating machine learning algorithms for fake news detection. In *2017 IEEE 15th student conference on research and development (SCOReD)*, 110–115. IEEE.

Guo, B.; Zhang, X.; Wang, Z.; Jiang, M.; Nie, J.; Ding, Y.; Yue, J.; and Wu, Y. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Han, T.; Kumar, A.; Agarwal, C.; and Lakkaraju, H. 2024. Medsafetybench: Evaluating and improving the medical safety of large language models. *Advances in Neural Information Processing Systems*, 37: 33423–33454.

Huang, T.; Yi, J.; Yu, P.; and Xu, X. 2025. Unmasking digital falsehoods: A comparative analysis of LLM-based misinformation detection strategies. In *2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*, 2470–2476. IEEE.

Jin, H.; Chen, R.; Zhou, A.; Zhang, Y.; and Wang, H. 2024a. Guard: Role-playing to generate natural-language jailbreakings to test guideline adherence of large language models. *arXiv preprint arXiv:2402.03299*.

Jin, H.; Hu, L.; Li, X.; Zhang, P.; Chen, C.; Zhuang, J.; and Wang, H. 2024b. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv preprint arXiv:2407.01599*.

Keskar, D.; Palwe, S.; and Gupta, A. 2020. Fake news classification on twitter using flume, n-gram analysis, and decision tree machine learning technique. In *Proceeding of International Conference on Computational Science and Applications: ICCSA 2019*, 139–147. Springer.

Kirch, N. M.; Field, S.; and Casper, S. 2024. What Features in Prompts Jailbreak LLMs? Investigating the Mechanisms Behind Attacks. *arXiv preprint arXiv:2411.03343*.

Krause, N. M.; Freiling, I.; Beets, B.; and Brossard, D. 2020. Fact-checking as risk communication: the multi-layered risk of misinformation in times of COVID-19. *Journal of Risk Research*, 23(7-8): 1052–1059.

Kumar, R.; Goddu, B.; Saha, S.; and Jatowt, A. 2024. Silver lining in the fake news cloud: Can large language models help detect misinformation? *IEEE Transactions on Artificial Intelligence*.

Lee, D.; Xie, S.; Rahman, S.; Pat, K.; Lee, D.; and Chen, Q. A. 2023. "Prompter Says": A Linguistic Approach to Understanding and Detecting Jailbreak Attacks Against Large-Language Models. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, 77–87.

Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; et al. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Li, X.; Zhou, Z.; Zhu, J.; Yao, J.; Liu, T.; and Han, B. 2023. Deepinception: Hypnotize large language model to be jailbreaker. *arXiv preprint arXiv:2311.03191*.

Liu, A.; Sheng, Q.; and Hu, X. 2024. Preventing and detecting misinformation generated by large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3001–3004.

Menz, B. D.; Kuderer, N. M.; Bacchi, S.; Modi, N. D.; Chin-Yee, B.; Hu, T.; Rickard, C.; Haseloff, M.; Vitry, A.; McKinnon, R. A.; et al. 2024. Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis. *BMJ*, 384.

Nakamura, K.; Levy, S.; and Wang, W. Y. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*.

Nathanson, S.; Yoo, Y.; Na, D.; Cao, Y.; and Watkins, L. 2024. A Step Towards Modern Disinformation Detection: Novel Methods for Detecting LLM-Generated Text. In *MILCOM 2024-2024 IEEE Military Communications Conference (MILCOM)*, 615–620. IEEE.

Pan, Y.; Pan, L.; Chen, W.; Nakov, P.; Kan, M.-Y.; and Wang, W. Y. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Ramesh, B.; Gursale, D.; Jopaul, A.; and Ernst, M. 2025. Reddit Misinformation Dataset. https://zenodo.org/records/14900167. To appear in DISMISS-FAKE'25: 1st Workshop on Disinformation and Misinformation in the Age of Generative AI, co-located with the 18th ACM WSDM, March 14, 2025, Hannover, Germany.

Rao, A. S.; Naik, A. R.; Vashistha, S.; Aditya, S.; and Choudhury, M. 2024. Tricking LLMs into Disobedience: Formalizing, Analyzing, and Detecting Jailbreaks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 16802–16830.

Russinovich, M.; Salem, A.; and Eldan, R. 2024. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. *arXiv preprint arXiv:2404.01833*.

Sager, M. A.; Kashyap, A. M.; Tamminga, M.; Ravoori, S.; Callison-Burch, C.; and Lipoff, J. B. 2021. Identifying and responding to health misinformation on reddit dermatology forums with artificially intelligent bots using natural language processing: design and evaluation study. *JMIR dermatology*, 4(2): e20975.

Saleh, H.; Alharbi, A.; and Alsamhi, S. H. 2021. OPCNN-FAKE: Optimized convolutional neural network for fake news detection. *IEEE Access*, 9: 129471–129489.

Scepanovic, S.; Martin-Lopez, E.; Quercia, D.; and Baykaner, K. 2020. Extracting medical entities from social media. In *Proceedings of the ACM conference on health, inference, and learning*, 170–181.

Shen, X.; Chen, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2024. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 1671–1685.

Sun, Y.; He, J.; Cui, L.; Lei, S.; and Lu, C.-T. 2024. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *arXiv preprint arXiv:2403.18249*.

Traore, I.; Woungang, I.; and Awad, A. 2018. *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*. Springer.

Wang, Z.; Cao, Y.; and Liu, P. 2024. Hidden You Malicious Goal Into Benign Narratives: Jailbreak Large Language Models through Logic Chain Injection. *arXiv preprint arXiv:2404.04849*.

Weidinger, L.; Mellor, J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Zeng, Y.; Lin, H.; Zhang, J.; Yang, D.; Jia, R.; and Shi, W. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14322–14350.

Zhang, H.; Lou, Q.; and Wang, Y. 2025. Towards Safe AI Clinicians: A Comprehensive Study on Large Language Model Jailbreaking in Healthcare. *arXiv preprint arXiv:2501.18632*.

Zhang, Y.; Sharma, K.; Du, L.; and Liu, Y. 2024. Toward mitigating misinformation and social media manipulation in llm era. In *Companion Proceedings of the ACM Web Conference 2024*, 1302–1305.

Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; and Deng, Y. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Zhou, J.; Zhang, Y.; Luo, Q.; Parker, A. G.; and De Choudhury, M. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–20.