

Article

Detecting Health Product Misinformation on Social Media Using Large Language Models Grounded in Biomedical Evidence

Sara Behnamian ^{1,*} , Zeinab Shahbazi ^{2,*} , Zahra Shahbazi ³  and Sadiqa Jafari ⁴ 

¹ Globe Institute, University of Copenhagen, 1353 Copenhagen, Denmark

² Research Environment of Computer Science (RECS), Kristianstad University, 291 88 Kristianstad, Sweden

³ Department of Environmental Engineering, University of Padova, 35131 Padova, Italy; zahra.shahbazi@studenti.unipd.it

⁴ School of Computing, Gachon University, Seongnam 13120, Republic of Korea; sadiqa1991@gachon.ac.kr

* Correspondence: sara.behnamian@sund.ku.dk (S.B.); zeinab.shahbazi@hkr.se (Z.S.)

Abstract

The spread of unverified health claims about drugs, dietary supplements, and alternative remedies on social media poses a growing public health concern. In this study, we present a retrieval-augmented generation (RAG) pipeline that uses large language models (LLMs) grounded in biomedical evidence from PubMed, openFDA adverse event reports, and NIH/NCCIH dietary supplement fact sheets to detect and classify health product misinformation. A total of 3493 health-related posts were collected from Reddit (948 posts across 12 subreddits) and YouTube (2545 video descriptions and comments), from which 8250 structured claims were extracted using Claude Haiku. Each claim was matched to biomedical evidence from three authoritative sources, achieving 79.4% evidence coverage, and classified into one of five veracity categories: supported (7.0%), unsupported (59.9%), exaggerated (22.4%), contradicted (2.0%), or dangerous (8.6%), together with an associated risk tier. Overall, 13.5% of claims were assigned high or critical risk. Cross-platform analysis showed that YouTube contained higher proportions of dangerous (11.3% vs. 2.9%) and exaggerated (27.0% vs. 12.4%) claims than Reddit. Compared with keyword-based and zero-shot transformer baselines, the LLM+RAG pipeline produced a more balanced and fine-grained classification of unsupported, exaggerated, contradicted, and dangerous claims. The most frequently implicated products were ashwagandha, kratom, black seed oil, turmeric, and ivermectin, with disease cure claims showing the highest dangerous classification rate (30.1%). These model-assigned results suggest that evidence-grounded LLM pipelines can support health misinformation surveillance, while also highlighting the need for expert validation and broader cross-platform evaluation.

Keywords: health misinformation; social media; retrieval-augmented generation (RAG); large language models (LLMs); biomedical evidence retrieval; veracity classification; risk assessment; natural language processing (NLP)



Academic Editors: Damianos P. Sakas and Nikolaos T. Giannakopoulos

Received: 16 April 2026

Revised: 10 May 2026

Accepted: 12 May 2026

Published: 14 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Health misinformation disseminated through digital media has become a major global public health concern. It can affect patient outcomes, public health management, and trust in healthcare institutions [1,2]. False or misleading claims may range from minor inaccuracies to harmful narratives. Such claims can contribute to vaccine hesitancy, delays in seeking medical care, and the use of unsafe or ineffective treatments [3,4]. Previous work

on health-related Reddit discussions, including topics such as obesity, diabetes, and hypertension, has also shown the persistence of false claims, including claims that herbal remedies can cure chronic diseases [5].

Traditional natural language processing (NLP) approaches to misinformation detection are often formulated as binary or multi-class text classification tasks. Although these methods can be useful for large-scale screening, they often rely on surface-level textual patterns and may have limited ability to capture context, implied meaning, and the strength of the underlying evidence [6–8]. This limitation is particularly important in health contexts, where a claim may be partially supported, exaggerated, unsupported, directly contradicted, or potentially dangerous depending on the available biomedical evidence.

Recent advances in large language models (LLMs) provide new opportunities for extracting and interpreting health-related claims from informal social media text. However, LLMs may also produce ungrounded or factually unreliable outputs when used without external evidence. Retrieval-augmented generation (RAG) offers a way to reduce this risk by linking model outputs to external, verifiable knowledge sources. In the health domain, this is especially important for claims about drugs, dietary supplements, and alternative remedies, where misleading statements may influence patient decisions and delay evidence-based care.

Despite these developments, current health misinformation detection systems still face three important limitations. First, many approaches classify posts directly, rather than extracting structured claims that can be checked against biomedical evidence. Second, many systems treat misinformation as a binary problem and do not distinguish between unsupported, exaggerated, contradicted, and dangerous claims. Third, relatively few studies examine how evidence-grounded misinformation patterns differ across social media platforms such as Reddit and YouTube.

To address these limitations, this study presents a RAG-based framework for detecting and classifying health product misinformation on social media. The pipeline combines LLM-based claim extraction, biomedical evidence retrieval from PubMed, openFDA adverse event reports, and NIH/NCCIH fact sheets, followed by risk-aware veracity classification. The system assigns each extracted claim to one of five categories: supported, unsupported, exaggerated, contradicted, or dangerous, together with an associated risk tier.

Problem Statement

Accurately identifying health misinformation on social media remains challenging because health-related posts are informal, heterogeneous, and often difficult to interpret without external evidence. Surface-level text classification alone is insufficient for this task, since the same product or treatment may be discussed in ways that are accurate, exaggerated, unsupported, or harmful depending on the specific claim being made. Moreover, the potential public health impact of a claim depends not only on whether it is factually correct, but also on the severity of the condition, the plausibility of harm, and the possibility that users may delay or replace evidence-based treatment.

This study therefore addresses the problem of evidence-grounded, risk-aware classification of health product claims extracted from social media. The goal is not only to detect whether misinformation is present, but also to distinguish different levels of evidential support and potential harm across Reddit and YouTube content.

The main contributions of this work are summarized as follows:

- A RAG-based framework that combines LLM-based claim extraction with biomedical evidence retrieval for health product misinformation detection.
- A multi-platform analysis of health-related claims collected from Reddit and YouTube.

- Structured extraction of claims involving products, asserted effects, and target conditions from informal social media text.
- Evidence matching against PubMed, openFDA adverse event reports, and NIH/NCCIH fact sheets.
- Five-class veracity classification combined with risk-tier assessment.
- Evaluation against keyword-based and zero-shot transformer baselines, including transformer models with and without retrieved evidence.
- Public release of code for reproducibility.

The remainder of this paper is organized as follows. Section 2 reviews related work on health misinformation detection, including traditional NLP approaches, transformer-based models, knowledge graph integration, privacy-preserving methods, and retrieval-augmented generation. Section 3 describes the proposed methodology, including data collection and preprocessing, LLM-based claim extraction, biomedical evidence retrieval, and veracity classification with risk assessment. Section 4 presents the experimental results across all pipeline stages, including comparison with keyword-based and transformer baselines and cross-platform analysis. Section 5 discusses the findings, platform differences, limitations, and implications. Section 6 concludes the paper and outlines directions for future work.

2. Literature Review

This section reviews prior work on health misinformation detection and highlights key methodological developments and remaining challenges. It covers traditional NLP methods, transformer-based models, knowledge graph integration, privacy-preserving approaches, and retrieval-augmented generation.

2.1. Health Misinformation and Impact

Health misinformation has evolved from isolated anecdotal content in early internet forums into a widespread public health issue amplified by modern social media platforms [9–11]. Such misinformation has been linked to vaccine hesitancy, reduced adherence to evidence-based medical practices, and the use of ineffective or unsafe treatments. Prior analyses of health-related Reddit discussions have also shown the persistence of false claims, including claims that herbal remedies can cure diabetes or that natural supplements can fully control hypertension. These examples illustrate how false health information can influence patient decision-making and healthcare use, underscoring the need for robust detection systems [3,12].

2.2. Health Fact-Checking

Automatic fact-checking has become increasingly important for addressing false health information on digital platforms. In the health domain, fact-checking aims to determine whether medical or health-related claims are accurate and to identify factual errors before they spread further [13]. Knowledge graphs, which display information in an organized manner called triples, have been used in earlier research to increase explanation clarity and accuracy [14]. For example, Cui et al. [15] created a model called DETERRENT that uses a graph attention network to direct article embeddings by utilizing extra information from a medical knowledge graph. Using triples from the medical knowledge graph, this model can identify pertinent knowledge facts and provide an explanation for the information's inaccuracy. LLMs such as GPT-4 provide an alternative approach because they can interpret complex language and generate natural-language explanations for fact-checking outputs. This can improve interpretability and accessibility. However, much LLM-based fact-checking research has focused on political or general-domain claims,

while medical and health-related claims remain less extensively studied. Although fact-checking accuracy has been assessed in a number of domains, including health-related content [16,17], the interpretability and reasoning processes that underlie model predictions in these settings have not been thoroughly examined. There are still issues with the dependability, safety, and factual accuracy of LLM-generated outputs in medical and public health applications [18], even though GPT-4 is said to have performed similarly to passing thresholds on the US Medical Licensing Exam [19].

2.3. Traditional and Transformer-Based NLP Approaches

Traditional NLP methods, including logistic regression, support vector machines (SVMs), TF-IDF features, and Word2Vec representations, formed the basis of early misinformation detection systems [20,21]. Although these methods established a foundation for automated text analysis, they have important limitations, including limited recall and difficulty capturing contextual and semantic relationships [22]. Transformer-based models such as BERT and GPT improved classification performance by providing stronger contextual representations [23]. However, these models can be computationally expensive and may still produce ungrounded or hallucinated outputs when used without external evidence. This limitation motivates the use of structured or retrieved knowledge sources to improve factual reliability, especially in high-stakes domains such as healthcare [24,25].

2.4. Knowledge Graph Integration for Misinformation Detection

Recent research has investigated the integration of knowledge graphs (KGs) with NLP models to improve factual grounding. KGs provide structured representations of domain-specific knowledge, allowing models to compare claims with established medical information [26]. For example, the DETERRENT framework uses KG-based representations obtained through Open Information Extraction (OpenIE) to support explainable healthcare misinformation detection. Despite these advances, integrating contextual language understanding with structured knowledge remains challenging, particularly at scale [15].

2.5. Privacy-Preserving Machine Learning in Healthcare

Privacy-preserving machine learning has also become important in digital health research because health-related data may contain sensitive personal information. Federated learning (FL) enables decentralized model training without directly exchanging raw data [27,28]. Other approaches, including differential privacy and secure multi-party computation (SMPC), aim to balance data utility with privacy protection [29,30]. Although these methods are not the focus of the present study, they are relevant to the broader deployment of health misinformation detection systems in privacy-sensitive settings [31].

2.6. LLMs, RAG, and Explainability

Recent research has highlighted the dual role of LLMs in misinformation detection: they can support health information analysis, but they can also generate or amplify inaccurate content [32,33]. Retrieval-augmented generation (RAG) addresses part of this problem by grounding model outputs in external, verifiable knowledge sources, thereby improving factual reliability [34]. Similar evidence-grounded LLM pipelines have also been developed outside the health domain; for example, Shahbazi and Behnamian proposed a climate misinformation detection and debunking system combining transformer-based classification, semantic similarity, stance detection, and RAG-based generation against authoritative climate-science sources [35]. Explainability techniques such as LIME, SHAP, and Integrated Gradients have also been used to improve model transparency [36]. Despite these developments, existing methods still lack integrated approaches that combine evidence retrieval, fine-grained veracity classification, and risk-aware assessment across

multiple social media platforms. Table 1 summarizes representative state-of-the-art studies on health misinformation detection.

Table 1. Summary of Representative Studies on Health Misinformation Detection.

Study	Methodology	Key Results	Limitations
Cui et al. [15]	Knowledge Graph + Graph Attention Network (DETERRENT framework)	Introduced explainable healthcare misinformation detection using structured medical knowledge	Requires curated knowledge graphs; limited scalability and domain transfer
Kumari et al. [37]	SCIBERT-based ensemble with domain-specific embeddings	Achieved F1 score of 0.715 on health fake-news datasets	Focused on news articles; moderate performance and limited generalization
Ayoub et al. [38]	DistilBERT + SHAP with data augmentation	Achieved high accuracy (0.972) and AUC (0.993) for COVID-19 misinformation detection	Domain-specific (COVID-19); relies on post-hoc explainability
Alghamdi et al. [39]	CT-BERT + BiGRU hybrid transformer model	Achieved state-of-the-art F1 score (98%) on COVID-19 fake news dataset	Limited to COVID-19 domain; lacks external knowledge grounding
Upadhyay et al. [40]	BERT with socio-contextual features	Improved detection accuracy by incorporating user and contextual metadata	Requires additional metadata; tested mainly on Twitter data
Men and Mariano [41]	BERT + BiLSTM + SHAP explainability	Provided interpretable predictions with improved classification performance	Post-hoc explanations; no grounding in verified biomedical sources
Khemani et al. [42]	Classical ML vs. Graph Convolutional Networks (GCN)	GCN achieved highest accuracy (93.86%) among tested models	Focus on classification; lacks explainability and evidence grounding
Padalko et al. [43]	BERT + BiLSTM + attention + topic modeling	Achieved high accuracy (99.47%) and extracted thematic misinformation patterns	Limited to specific datasets; lacks external evidence verification
Gargari et al. [44]	Retrieval-Augmented Generation (RAG) with semantic matching	Improved factual accuracy and retrieval performance in health information tasks	Focuses on retrieval rather than full misinformation classification

3. Materials and Methods

This study proposes a retrieval-augmented generation (RAG)-based framework for detecting and classifying health product misinformation from social media content. The pipeline consists of four stages: (i) data collection and preprocessing from Reddit and YouTube, (ii) LLM-based claim extraction and structuring, (iii) biomedical evidence retrieval from PubMed, openFDA adverse event reports, and NIH/NCCIH fact sheets, and (iv) veracity classification with risk-tier assessment.

The framework is designed to move beyond direct post-level classification. Instead, each social media post is first converted into one or more structured claims, including the product, asserted effect, target condition, claim strength, and risk category. Each claim is then linked to external biomedical evidence before classification. This design allows the system to distinguish between claims that are supported, unsupported, exaggerated, contradicted, or dangerous, rather than treating misinformation as a binary label. Figure 1 illustrates the overall architecture of the proposed system.

The pipeline begins with keyword-guided collection of health-related content from Reddit and YouTube. After preprocessing, an LLM extracts structured claims from each post. These claims are then used to retrieve biomedical evidence from PubMed, openFDA, and NIH/NCCIH sources. Finally, the LLM evaluates each claim against the retrieved evidence and assigns a veracity label, confidence score, reasoning text, and risk tier. The resulting classifications are compared with a keyword-based baseline and two zero-shot transformer baselines, one without retrieved evidence and one with retrieved evidence.

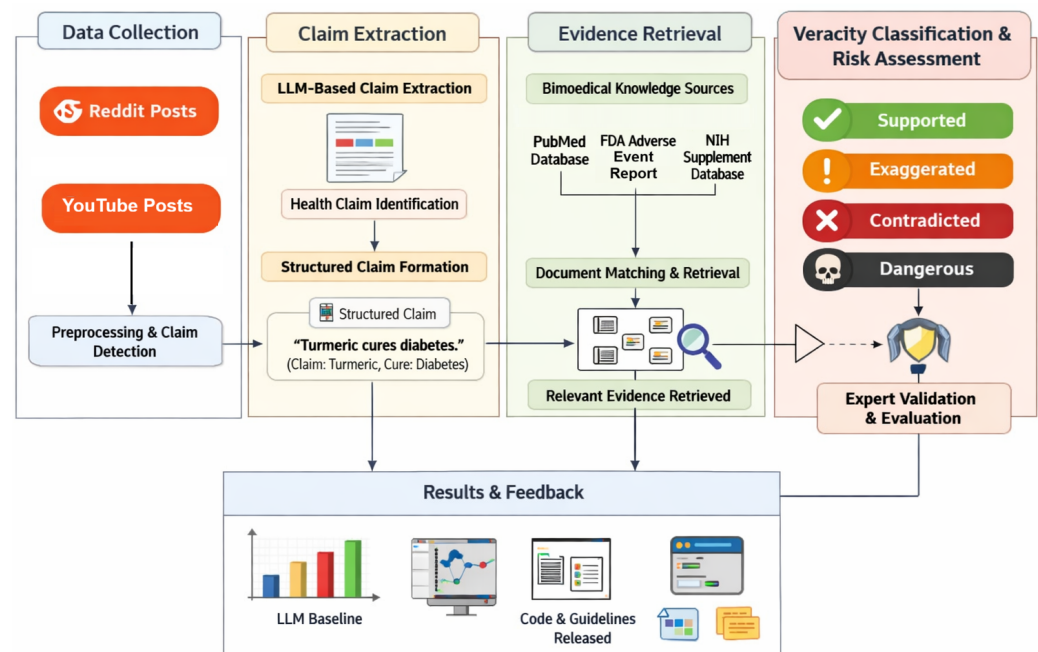


Figure 1. Architecture of the proposed RAG-based health misinformation detection framework, combining social media data collection, LLM-based claim extraction, biomedical evidence retrieval, and veracity classification with risk assessment.

3.1. Data Collection and Preprocessing

The first stage collected health-related content from Reddit and YouTube. These two platforms were selected because they cover two complementary modes of online health discourse: long-form text-based discussion in topic-specific communities (Reddit), and video-driven content with high-volume comment threads (YouTube). Both also expose public, programmatically accessible content endpoints—Reddit’s public JSON endpoints and the YouTube Data API v3—that allow large-scale, reproducible data collection without scraping, which was not the case for some other major platforms (e.g., TikTok, Instagram, X) at the time of data collection. Reddit posts and top-level comments were collected from targeted health-related subreddits using public JSON endpoints. YouTube video descriptions and comments were collected using the YouTube Data API v3. Both collection scripts used predefined health-product and treatment-related keywords to identify potentially relevant content.

The Reddit collection included subreddits related to supplements, alternative health, nootropics, herbalism, nutrition, biohacking, skincare, hormonal health, chronic conditions, cancer, weight loss, fasting, and ketogenic diets. The YouTube collection used 20 health-related search queries, including queries related to turmeric, natural remedies, ivermectin, black seed oil, detox claims, ashwagandha, CBD oil, kratom, hydrogen peroxide therapy, and supplement reviews.

Collected entries were filtered to retain health-related English-language content and remove duplicate or near-duplicate records. Additional preprocessing included basic text cleaning and standardization for downstream processing. Each retained entry was stored with metadata, including platform, source identifier, timestamp, URL, author information where available, and engagement-related fields such as score, likes, or comment counts.

The resulting dataset should be interpreted as a keyword-guided sample of public social media discourse rather than a fully representative sample of all health-related content online. The data collection and preprocessing pipeline is illustrated in Figure 2.

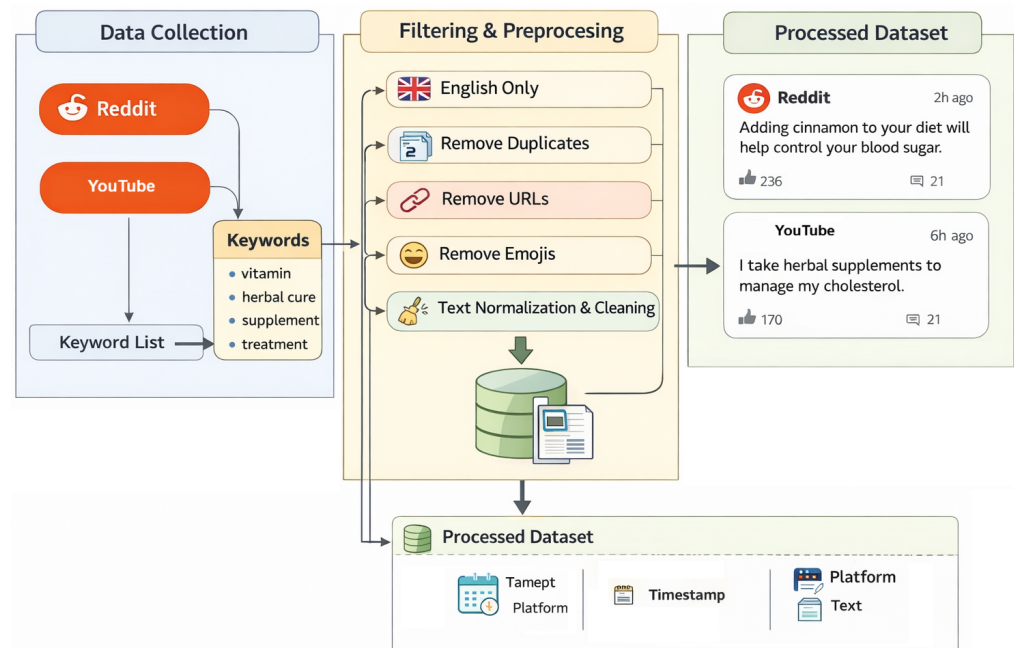


Figure 2. Overview of the data collection and preprocessing pipeline. Health-related content is collected from Reddit and YouTube using keyword-guided queries. The collected data undergo filtering, duplicate removal, and basic text standardization to produce a structured dataset for downstream claim extraction.

3.2. LLM-Based Claim Extraction and Structuring

The second stage extracted structured health claims from the collected social media posts using Claude Haiku via the Anthropic API. Each post was passed to the LLM with a structured prompt asking it to identify all health-related claims and return valid JSON. For each extracted claim, the model recorded the product or substance, the claimed effect, the target condition, the claim strength, a verbatim quote, and a risk category.

The claim strength label was restricted to one of three categories: definitive, suggestive, or anecdotal. The risk category was restricted to one of eight categories: supplement efficacy, drug misuse, disease cure, weight loss, detox, anti-aging, mental health, or other. Posts with no identifiable health claim were returned as empty claim lists and excluded from downstream evidence retrieval.

This structured representation allowed the pipeline to move from noisy social media text to claim-level verification. For example, a statement such as “Turmeric cures diabetes” would be represented by the product “turmeric”, the claimed effect “cures diabetes”, and the target condition “diabetes”. The LLM-based claim extraction and structuring process is illustrated in Figure 3.

3.3. Biomedical Evidence Retrieval

The third stage retrieved biomedical evidence for each extracted claim. Evidence retrieval used three sources: PubMed/MEDLINE through the NCBI Entrez API, openFDA adverse event reports, and NIH Office of Dietary Supplements or NCCIH fact-sheet references.

For PubMed retrieval, a query was constructed from the structured claim fields, primarily the product name, target condition, and claimed effect. The PubMed search prioritized higher-quality evidence by filtering for systematic reviews, meta-analyses, and randomized controlled trials. For each claim, the script retrieved article metadata, including PMID, title, abstract, journal, year, publication type, and PubMed URL.

For safety-related evidence, the pipeline queried the openFDA drug adverse event API using the product name. Retrieved records included reported adverse reactions, seriousness

indicators, and report dates. For common dietary supplements and alternative remedies, the pipeline also checked whether an NIH or NCCIH fact sheet was available using a curated product-name dictionary.

The output of this stage was a claim–evidence record containing the original structured claim together with any retrieved PubMed articles, openFDA adverse event records, and NIH/NCCIH references. This evidence record was then passed to the veracity classification stage. The biomedical evidence retrieval process is illustrated in Figure 4.

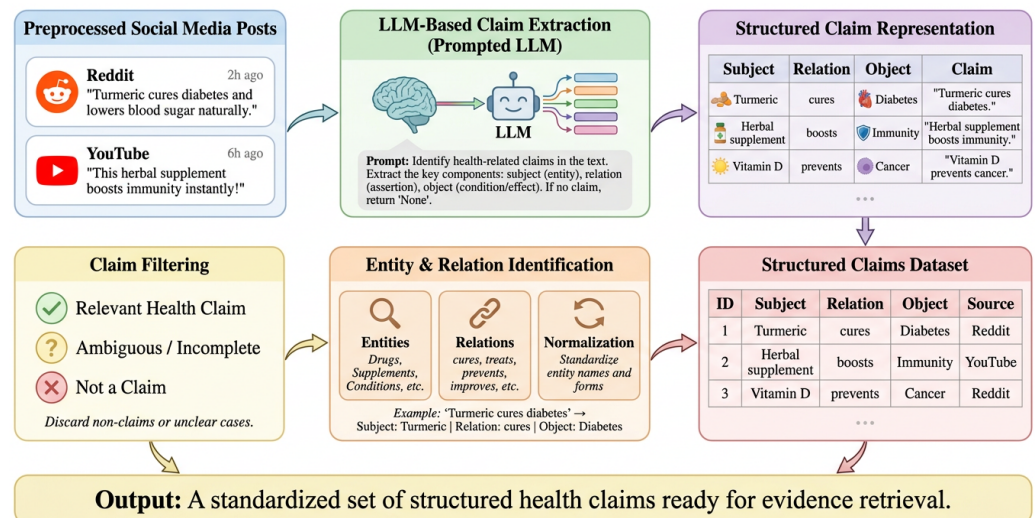


Figure 3. Overview of the LLM-based claim extraction and structuring pipeline, transforming unstructured social media posts into standardized claim representations for downstream verification.

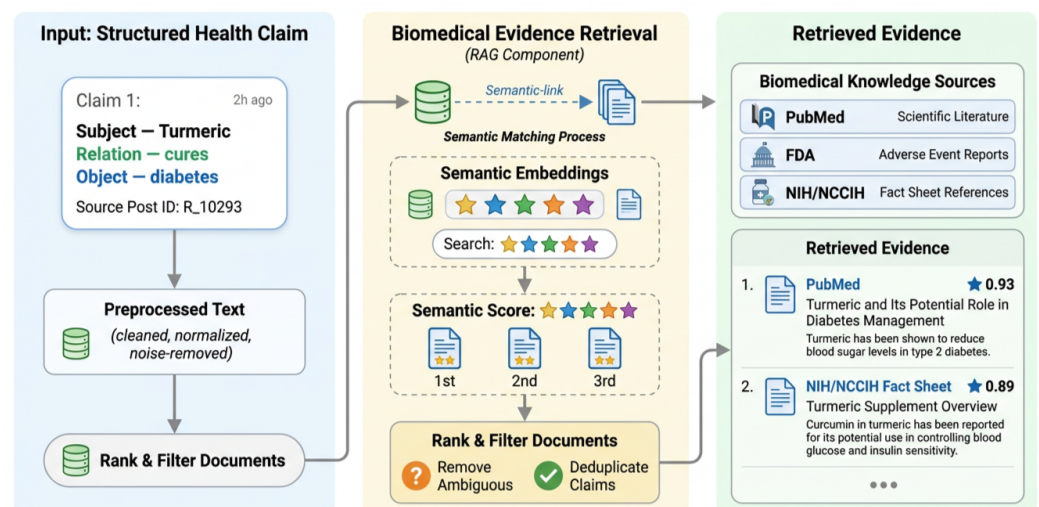


Figure 4. Biomedical evidence retrieval pipeline. Structured claims are used to retrieve evidence from PubMed, openFDA adverse event reports, and NIH/NCCIH fact-sheet references before veracity classification.

3.4. Veracity Classification and Risk Assessment

In the final stage, each claim was classified using an LLM-based RAG prompt. The prompt included the structured claim fields and the retrieved biomedical evidence. The model was instructed to assign one of five veracity labels: supported, unsupported, exaggerated, contradicted, or dangerous. It also assigned a confidence score, a risk tier, a brief reasoning statement, key evidence, and a recommendation.

The risk tier was restricted to four levels: low, moderate, high, and critical. Low-risk claims were defined as unlikely to cause direct harm, whereas moderate-risk claims

could waste money or delay appropriate care. High-risk claims could cause direct harm, and critical-risk claims involved potentially severe harm, especially for vulnerable populations or serious conditions.

For comparison, the study included three baselines. The first was a keyword-based classifier using curated trigger phrases for dangerous and exaggerated claims, with unsupported as the default label. The second was a zero-shot transformer baseline using facebook/bart-large-mnli applied to the claim text without retrieved evidence. The third used the same transformer model with a compact summary of retrieved evidence added to the input. These baselines were used to assess the contribution of evidence retrieval and evidence-aware LLM reasoning.

4. Results

This section presents the experimental results obtained at each stage of the proposed pipeline. Section 4.1 reports the data collection statistics across Reddit and YouTube, including post counts, temporal coverage, and platform-level characteristics. Section 4.2 summarizes the LLM-based claim extraction outcomes, including the number of structured claims, their distribution across risk categories and claim strengths, and the most frequently mentioned products and conditions. Section 4.3 describes the biomedical evidence retrieval results from PubMed, openFDA, and NIH/NCCIH fact sheets, including coverage statistics and source overlap. Section 4.4 presents the veracity classification results from the LLM+RAG pipeline and compares them against keyword-based and zero-shot transformer baselines, with cross-platform and category-level analyses. Finally, Section 4.5 provides illustrative classification examples for each veracity category.

4.1. Data Collection

A total of 3493 unique posts were collected across two platforms: 948 from Reddit (27.1%) and 2545 from YouTube (72.9%). Reddit data were gathered from 14 targeted health-related subreddits using public JSON endpoints, of which 12 returned results. The highest-yielding subreddits focused on supplements, hormonal health, nootropics, biohacking, and ketogenic diets, contributing between 88 and 175 posts each. Of the 948 Reddit entries, 814 were original posts and 134 were top-level comments collected from high-scoring threads (score > 5). Reddit posts spanned from July 2024 to April 2026, with a mean text length of 1202 characters (median 818), a mean score of 260.9 (median 16), and a mean of 72.0 comments per post.

YouTube data were collected via the YouTube Data API v3 using 20 health-related search queries, retrieving up to 10 videos per query and up to 50 comments per video. The resulting 2545 entries comprised 193 video descriptions and 2352 user comments drawn from 198 unique videos and 141 unique channels. The most frequently represented channels were health-focused practitioner accounts, with the top three channels contributing 13, 7, and 6 videos, respectively. YouTube content spanned from December 2010 to April 2026, reflecting the platform's longer content lifespan. A total of 2292 unique comment authors contributed to the YouTube dataset. YouTube entries were substantially shorter than Reddit posts, with a mean text length of 299 characters (median 210), consistent with the comment-dominant composition of this subset. All posts were deduplicated by unique identifier, and no duplicate entries were present in the final dataset.

4.2. Claim Extraction

A total of 8250 structured health claims were extracted from the 3493 collected posts using Claude Haiku (claude-haiku-4-5-20251001) via the Anthropic API. Of the input posts, 3219 (92.2%) yielded at least one claim, while 274 posts (7.8%) contained no identifiable

health claims and were excluded from further analysis. The mean number of claims per post was 2.56 (median 2), with a range of 1 to 17. The majority of posts produced one to three claims: 778 posts yielded a single claim, 1219 yielded two, and 632 yielded three, while 19 posts produced 11 or more claims.

By platform, YouTube contributed 5646 claims (68.4%) and Reddit contributed 2604 (31.6%). Among YouTube claims, 5339 originated from user comments and 307 from video descriptions. On Reddit, the highest-yielding subreddits in terms of extracted claims focused on supplements, nootropics, biohacking, hormonal health, and ketogenic diets, contributing between 253 and 524 claims each.

Each extracted claim was assigned a risk category and a claim strength label by the LLM. The most prevalent risk category was supplement efficacy, accounting for 4478 claims (54.3%), followed by other (944; 11.4%), drug misuse (748; 9.1%), disease cure (674; 8.2%), mental health (563; 6.8%), weight loss (442; 5.4%), detox (297; 3.6%), and anti-aging (104; 1.3%). Regarding claim strength, 3507 claims (42.5%) were classified as suggestive, 3271 (39.6%) as anecdotal, and 1469 (17.8%) as definitive. YouTube exhibited a higher proportion of definitive claims (19.3%) compared to Reddit (14.5%), while Reddit had a slightly higher proportion of anecdotal claims (43.4% vs. 37.9%).

The extraction identified 2653 unique products and 1619 unique target conditions across the dataset. The most frequently mentioned products were ashwagandha (417 claims), kratom (402), black seed oil (371), turmeric (259), ivermectin (242), vitamin D (222), and CBD oil (210). The most common target conditions were general health (2826 claims), followed by COVID-19 (224), anxiety (218), cancer (202), high cholesterol (168), common cold (150), and depression (124). Platform-level differences were also observed: YouTube claims were more concentrated in the disease cure (10.8% vs. 2.4%) and drug misuse (10.8% vs. 5.4%) categories, whereas Reddit claims had a higher proportion in the other category (20.0% vs. 7.5%).

4.3. Evidence Retrieval

Biomedical evidence was retrieved for all 8250 extracted claims using three sources: PubMed (via the NCBI Entrez API), the openFDA adverse event reporting system, and the NIH Office of Dietary Supplements and NCCIH fact sheets. Of the 8250 claims, 6553 (79.4%) were matched to at least one evidence source, while 1697 (20.6%) returned no evidence from any of the three databases. The mean number of evidence items per claim was 4.11 (median 4), with a maximum of 9.

PubMed returned results for 3539 claims (42.9%), yielding a total of 14,427 article records corresponding to 5208 unique publications (PMIDs). Among claims with at least one PubMed result, the mean number of articles retrieved was 4.08. The retrieved literature was dominated by high-quality evidence types: systematic reviews (8379 records), randomized controlled trials (5327), meta-analyses (4747), and network meta-analyses (481), consistent with the publication-type filter applied during the search. The openFDA adverse event database returned results for 5600 claims (67.9%), yielding 16,236 adverse event records in total. NIH or NCCIH fact sheet references were matched to 3210 claims (38.9%), based on a curated dictionary of 19 common supplement names.

Regarding source overlap, 2957 claims were matched by both PubMed and openFDA sources, 582 had PubMed evidence only, 2643 had openFDA evidence only, and 1513 claims were matched by all three sources simultaneously. Evidence coverage varied by platform: YouTube claims had a higher mean evidence count (4.44 vs. 3.37 for Reddit) and a lower proportion of zero-evidence claims (16.3% vs. 29.8%). YouTube claims were also more frequently matched to NIH references (48.3% vs. 18.5%), reflecting the higher prevalence of well-known supplement names in YouTube content.

Evidence coverage also differed across risk categories. Drug misuse and mental health claims had the highest mean evidence counts (5.4 each) and the lowest proportions of zero-evidence claims (8.7% and 13.0%, respectively). Supplement efficacy claims, which constituted the largest category, had a mean of 4.3 evidence items, with 14.0% returning no results. In contrast, detox claims had substantially lower coverage, with a mean of 0.9 evidence items and 69.4% returning no results, likely reflecting the limited representation of detoxification products in the biomedical literature. Weight loss (35.5% zero-evidence) and anti-aging (38.5%) claims also showed relatively lower evidence availability.

4.4. Veracity Classification

All 8250 claims were classified using the LLM+RAG pipeline, in which Claude Haiku evaluated each claim in relation to the retrieved biomedical evidence, and compared with a keyword-based baseline. The LLM+RAG classifier assigned each claim a veracity label, a risk tier, a confidence score, and a textual reasoning.

The most frequent veracity label was unsupported (4945 claims; 59.9%), indicating that the model could not establish a clear evidential basis for the claim. Because this label can arise from two distinct situations—claims for which no biomedical evidence was retrieved, and claims for which evidence was retrieved but did not substantiate the asserted effect—we report a sub-breakdown of the unsupported category. Of the 4945 unsupported claims, 2438 (49.3%) returned no evidence from any of the three databases (“no-evidence unsupported”), and 2507 (50.7%) had at least one retrieved record but no record provided clear support for the claim (“evidence-retrieved unsupported”). The two sub-types correspond to qualitatively different epistemic situations: the first reflects a coverage gap of the retrieval sources for the relevant product or claim, whereas the second reflects evidence that exists but does not align with the claim as stated. We discuss the implications of this distinction in Section 5.5. Exaggerated claims, where partial evidence existed but the claim overstated the effect, accounted for 1851 claims (22.4%). A total of 713 claims (8.6%) were classified as dangerous, meaning the claim contradicted safety data or promoted potentially harmful behavior. Supported claims, confirmed by high-quality evidence, comprised 576 (7.0%), and contradicted claims, directly refuted by evidence, comprised 165 (2.0%). The overall veracity distribution is shown in Figure 5.

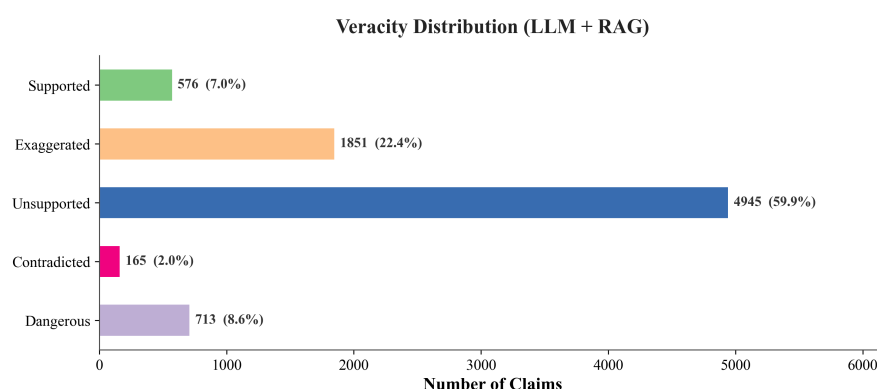


Figure 5. Veracity distribution of 8250 health claims classified by the LLM+RAG pipeline. The majority of claims (59.9%) were unsupported by biomedical evidence, while 22.4% were exaggerated and 8.6% were classified as dangerous.

The risk tier distribution reflected the predominance of moderate-harm claims: 4706 claims (57.0%) were assigned the moderate tier, 2431 (29.5%) were low, 539 (6.5%) were high, and 574 (7.0%) were critical. Taken together, 1113 claims (13.5%) fell into the high or critical risk tiers (Figure 6). The mean classification confidence across all claims

was 0.87 (median 0.85, standard deviation 0.08), with confidence scores ranging from 0.15 to 0.99. Confidence was highest for dangerous claims (mean 0.94, median 0.95) and contradicted claims (mean 0.90, median 0.92), and lowest for supported claims (mean 0.81, median 0.82), suggesting that the model expressed greater certainty when identifying harmful or refuted content.

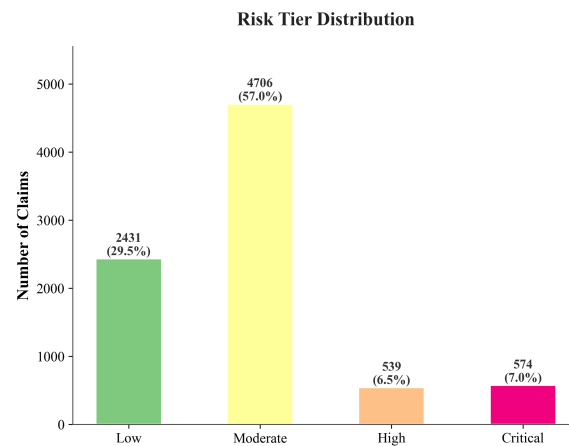


Figure 6. Distribution of risk tiers assigned by the LLM+RAG classifier. A total of 13.5% of claims were assigned high or critical risk.

Compared with the keyword baseline, the LLM+RAG approach assigned claims to a broader and more informative set of veracity categories. The baseline classified 96.8% of all claims as unsupported, detecting only 131 exaggerated and 133 dangerous claims. In contrast, the LLM+RAG pipeline identified 14.1 times more exaggerated claims and 5.4 times more dangerous claims than the baseline. Overall agreement between the two methods was 60.1%. Among the 7986 claims labeled unsupported by the baseline, the LLM+RAG pipeline reclassified 1763 as exaggerated, 622 as dangerous, 575 as supported, and 163 as contradicted, demonstrating the added value of evidence-grounded reasoning over surface-level keyword matching.

Ablation Against Pretrained Transformer Baselines

To address the concern that a keyword-only baseline is insufficient to demonstrate the contribution of the proposed pipeline, we conducted a controlled ablation against two additional baselines based on the same pretrained transformer model, facebook/bart-large-mnli [45], a 407M-parameter natural-language-inference model. Both transformer baselines were used in zero-shot mode, with each claim scored against five hypothesis statements (one per veracity class) and assigned the highest-scoring label. The two transformer baselines differ only in whether they are given access to the retrieved biomedical evidence:

- **Keyword baseline** (no language model, no evidence)—regular-expression matching against curated trigger phrases, with an UNSUPPORTED default for unmatched claims.
- **Transformer (no RAG)**—BART-MNLI applied to the claim text only (verbatim quote, product, asserted effect), without access to retrieved evidence.
- **Transformer (with RAG)**—BART-MNLI applied to the same claim text plus a compact summary of the retrieved evidence used by the full pipeline (top PubMed article title and 400-character abstract excerpt, top six FDA adverse-reaction terms, and an NIH fact-sheet flag), concatenated within the model’s 1024-token input window.

This three-baseline design isolates the contributions of (i) evidence retrieval and (ii) language-model reasoning capacity. The keyword baseline isolates the joint effect of removing both components; the no-RAG transformer isolates the effect of removing

only evidence retrieval; the with-RAG transformer isolates the effect of substituting the high-capacity LLM with a smaller NLI model while keeping evidence access. Each system was evaluated against the LLM + RAG labels on the full set of 8250 claims, treating those labels as the reference distribution for the purpose of the ablation. Results are summarized in Table 2.

Table 2. Three-baseline ablation against the full LLM + RAG pipeline. All systems are evaluated against the LLM + RAG labels on the full 8250-claim dataset. The transformer baselines use facebook/bart-large-mnli (zero-shot, no fine-tuning); they differ only in whether the retrieved biomedical evidence is included in the input. Per-class precision (P), recall (R), and F1 are reported with respect to the LLM + RAG distribution.

Class (n in LLM+RAG)	Metric	Keyword	BART, No RAG	BART + RAG
Overall accuracy		—	0.217	0.476
Macro F1		—	0.131	0.225
Supported (576)	P	—	0.085	0.160
	R	—	0.047	0.932
	F1	—	0.060	0.273
Unsupported (4945)	P	—	0.662	0.766
	R	—	0.021	0.676
	F1	—	0.040	0.718
Exaggerated (1851)	P	—	0.220	0.226
	R	—	0.833	0.004
	F1	—	0.348	0.007
Contradicted (165)	P	—	0.103	0.062
	R	—	0.024	0.158
	F1	—	0.039	0.089
Dangerous (713)	P	—	0.166	0.203
	R	—	0.168	0.021
	F1	—	0.167	0.038
Default collapse		UNSUP. 96.8%	EXAGG. 85.1%	UNSUP. + SUPP. 93.6%

Figure 7 compares the veracity classifications produced by the LLM + RAG pipeline and the keyword-based baseline.

The keyword baseline assigned the default UNSUPPORTED label to 96.8% of claims, reflecting its inability to detect overstatement, contradiction, or harm patterns through surface keyword matching alone. Removing this restriction by introducing a transformer language model (BART zero-shot, no RAG) did not improve performance: instead, the model collapsed 85.1% of claims (7017 of 8250) into the EXAGGERATED class, achieving an accuracy of 21.7% and a macro F1 of 0.131 against the LLM + RAG labels. Without access to evidence, the transformer recovered only 2.1% of unsupported claims and 4.7% of supported claims—the categories for which evidence retrieval is most informative. Adding retrieved evidence to the transformer’s input (BART + RAG) approximately doubled both metrics, raising accuracy to 47.6% and macro F1 to 0.225. The improvement was concentrated in the UNSUPPORTED class, where F1 rose from 0.040 to 0.718 (recall from 0.021 to 0.676), and in the SUPPORTED class, where recall reached 0.932 although precision remained low (0.160) due to over-prediction. However, the transformer-with-RAG system still failed on the categories that require finer-grained reasoning over the retrieved evidence: EXAGGERATED F1 fell to 0.007 (only 31 of 1851 LLM+RAG-flagged claims recovered), DANGEROUS F1 to 0.038 (74 of 713), and CONTRADICTED F1 to 0.089 (26 of 165).

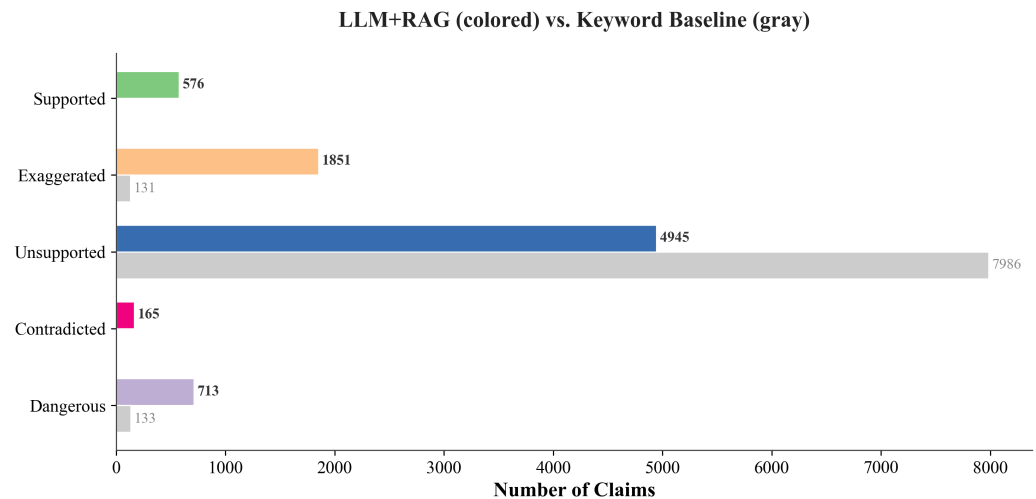


Figure 7. Comparison of veracity classifications between the LLM+RAG pipeline (colored) and the keyword-based baseline (gray). The baseline classified 96.8% of claims as unsupported, while the LLM+RAG approach detected 14.1 times more exaggerated and 5.4 times more dangerous claims.

Taken together, the three baselines bracket the failure modes of evidence-grounded misinformation detection. Without language modeling and without evidence (keyword baseline), the system cannot make any classification decision and defaults to UNSUPPORTED. With language modeling but no evidence (transformer no-RAG), the model over-assigned the EXAGGERATED label in the absence of grounding. With language modeling and evidence access but limited reasoning capacity (transformer + RAG), it can detect the presence or absence of supporting evidence (UNSUPPORTED F1 = 0.718) but not the misrepresentation of evidence (EXAGGERATED, DANGEROUS, CONTRADICTED F1 all < 0.10). The full LLM+RAG pipeline achieves the more even distribution shown in Section 4.4 only because both evidence retrieval and a language model capable of reasoning over that evidence are present. The ablation therefore supports the design hypothesis that evidence retrieval and LLM-based reasoning are jointly necessary for fine-grained misinformation classification.

Cross-platform analysis revealed substantial differences in misinformation profiles (Figure 8). YouTube exhibited a markedly higher proportion of exaggerated claims (27.0% vs. 12.4% for Reddit) and dangerous claims (11.3% vs. 2.9%), while Reddit had a higher proportion of unsupported claims (74.9% vs. 53.0%). The proportion of claims assigned high or critical risk was also considerably higher on YouTube (17.7%) than on Reddit (4.4%).

Veracity patterns also varied across claim categories (Figure 9). Disease cure claims had the highest rate of dangerous classification (30.1%) and exaggeration (38.7%), with only 1.3% classified as supported. Drug misuse claims showed similarly elevated rates of dangerous (18.3%) and exaggerated (31.4%) labels. In contrast, supplement efficacy claims, the largest category, had a lower dangerous rate (5.2%) but a substantial exaggeration rate (21.0%). Detox claims were notable for having zero supported claims and a high unsupported rate (86.2%), consistent with the limited evidence coverage observed in the retrieval stage.

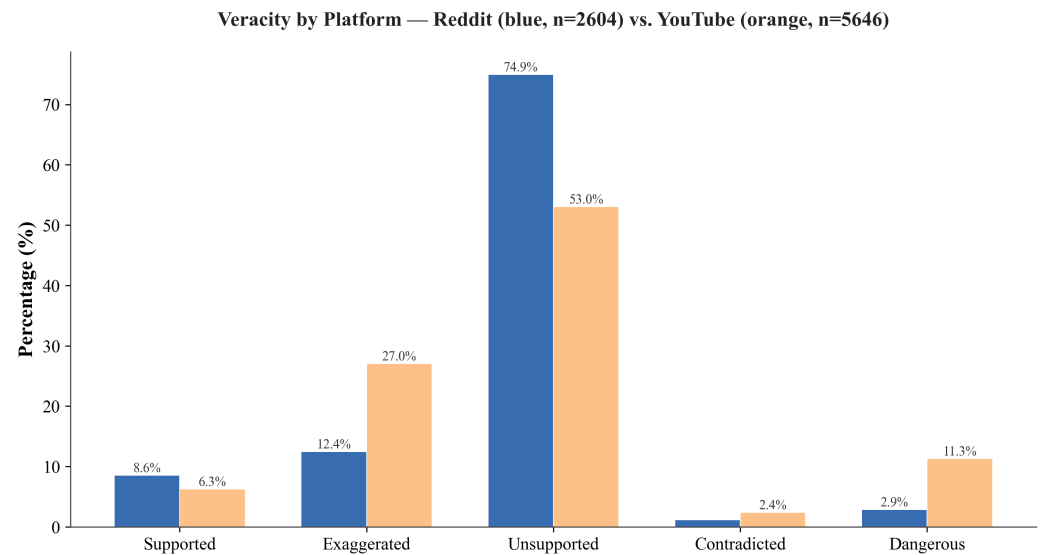


Figure 8. Veracity distribution by platform (percentage). YouTube exhibited higher rates of exaggerated (27.0% vs. 12.4%) and dangerous (11.3% vs. 2.9%) claims compared to Reddit.

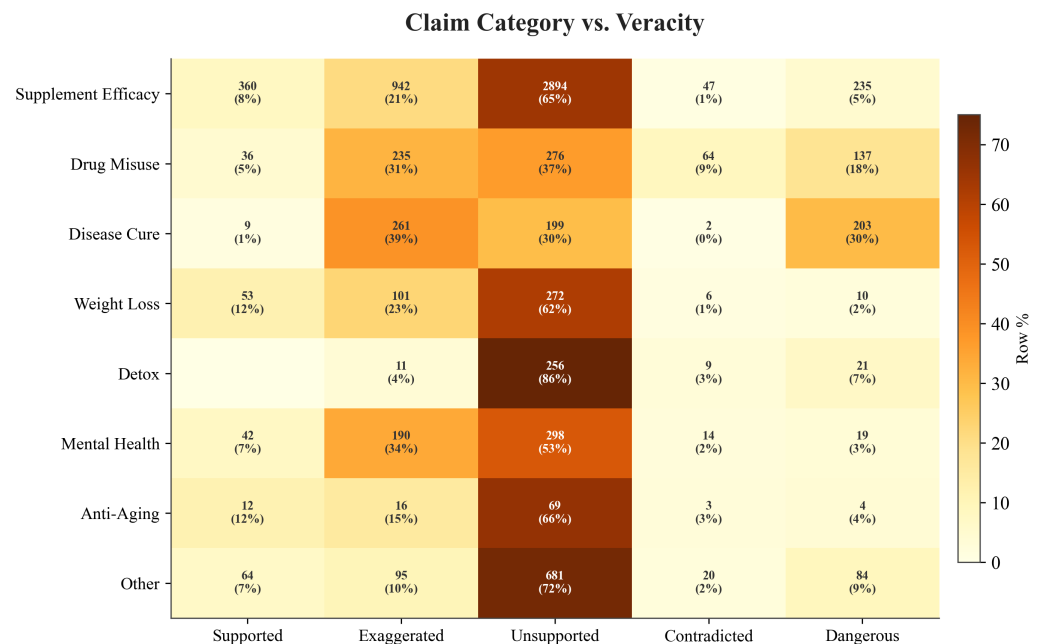


Figure 9. Row-normalized heatmap of claim category versus veracity classification. Disease cure claims had the highest dangerous rate (30%), while detox claims were 86% unsupported.

4.5. Illustrative Classification Examples

To illustrate the pipeline’s classification behavior, this section presents one representative example for each veracity category, selected from the highest-confidence classifications. For each example we report the verbatim claim text, the retrieved evidence (number of PubMed articles, openFDA records, and NIH or NCCIH references), and a summary of the model’s evidence-grounded reasoning, allowing the reader to inspect the link between retrieved sources and the assigned label.

As an example of a **supported** claim, a Reddit post stating that vitamin D3 supplementation increases blood vitamin D levels was classified with 0.95 confidence and assigned a low risk tier. The system retrieved multiple meta-analyses and systematic reviews confirming a well-established dose–response relationship, consistent with NIH fact sheet guidance. The verbatim claim was: “This is what i got while taking 5000 IU a day, i decided that

since I'm only at 49 to start doing 10,000 IU's a day". Five PubMed articles, three openFDA adverse-event records, and a matching NIH fact sheet were retrieved. The model identified two of the PubMed records as systematic reviews and meta-analyses specifically examining vitamin D supplementation's ability to raise serum 25(OH)D levels, which directly support the core mechanism of the claim. The low risk tier reflects the well-characterized safety profile of vitamin D3 at the dosages discussed and the absence of contradicting safety signals in the retrieved evidence.

An **exaggerated** claim from Reddit asserted that mesenchymal stem cells (MSCs) can completely cure Alzheimer's disease (confidence 0.95, high risk). While the retrieved PubMed evidence included studies showing potential therapeutic effects in animal models, no human trials supported a definitive cure claim; the literature characterized MSC therapy as "promising" rather than curative. The verbatim claim was the post title "1.5 M Stem Cells Completely Curing Alzheimer's Disease". The model retrieved four PubMed articles and zero openFDA records; one meta-analysis examined MSC effects on cognitive deficits in animal models rather than humans, and another characterized MSC therapy as a "promising treatment" rather than a cure. The exaggerated label thus reflects partial evidential support for a therapeutic effect alongside an overstated magnitude ("completely curing"), and the high risk tier captures the potential harm of patients with a serious neurodegenerative condition acting on a definitive cure claim that the evidence does not support.

Among **unsupported** claims, a YouTube comment attributed environmental benefits to kratom, specifically that it reduces riverbank erosion and abrasion (confidence 0.98, low risk). All retrieved evidence addressed kratom's pharmacological properties in humans, with zero results pertaining to environmental or ecological applications. The verbatim claim was: "kratom can become the main foundation to reduce the impact of erosion and abrasion on the riverbank". The model retrieved one PubMed article, three openFDA records, and an NIH reference, all addressing kratom's pharmacology and adverse health events in humans, with none addressing environmental or geomorphological applications. The unsupported label was therefore assigned not because the claim contradicts the evidence, but because no evidence in the biomedical retrieval scope is relevant to this domain. The low risk tier reflects that the claim concerns environmental impact rather than a directly health-relevant assertion.

A claim from YouTube stated that ivermectin had been approved by the NIH for COVID-19 treatment (confidence 0.98, high risk). The retrieved evidence included the Cochrane 2022 systematic review, which reported that evidence for ivermectin's efficacy against COVID-19 remains conflicting, and no source confirmed any NIH approval for this indication. The verbatim claim was: "ivermectin was just approved by the NIH for covid about 10 days ago". The model retrieved five PubMed articles and three openFDA records. The Cochrane 2022 review explicitly stated that "evidence on ivermectin for prevention of SARS-CoV-2 infection and COVID-19 treatment is conflicting", directly contradicting the definitive approval claim, and no retrieved source contained any NIH approval statement. The high risk tier reflects the public-health implications of misattributing regulatory endorsement to a treatment whose efficacy remains unestablished.

Finally, a **dangerous** claim from Reddit described using apple-flavored livestock dewormer paste to treat gastrointestinal symptoms in humans (confidence 0.99, critical risk). The system flagged this as dangerous because the product is formulated and regulated exclusively for veterinary use, lacks human safety and dosing data, and carries risks of unknown toxicity. No FDA approval for human use was found in any retrieved source. The verbatim claim was: "After just one dose I didn't wake up at 4:30. That was the first big shift." The model retrieved zero records from PubMed, openFDA, and NIH for this product's human use, which itself is informative: the absence of any evidence base

for a substance being self-administered as medicine, combined with its veterinary-only regulatory status, was the basis for the dangerous classification. The critical risk tier reflects the unknown human dosing, lack of quality control, and the potential for serious adverse events from off-label veterinary product use.

These examples demonstrate the pipeline's capacity to distinguish between varying levels of evidential support, from well-documented supplement effects to dangerous off-label use of veterinary products, using grounded biomedical evidence rather than surface-level text features alone.

The five examples above are all drawn from a single platform per class. To verify that the pipeline behaves consistently across different platforms and across the full risk-tier range, we present five further examples in which each veracity class is illustrated on the alternate platform (Reddit if the original was on YouTube, and vice versa) and, where possible, at a different risk tier.

For **supported** claims, a Reddit post about a different therapeutic class—low-dose isotretinoin (Accutane) for acne—was classified with 0.95 confidence and a low risk tier. The verbatim claim was: “They also prescribed me low doses (10 mg?) of isotretinoin (accutane) which completely cleared my acne”. Five PubMed articles were retrieved, of which one explicitly stated that “oral isotretinoin is the most effective treatment for moderate to severe acne” and confirmed that low-dose regimens (0.5 mg/kg/day) achieve effectiveness comparable to higher doses. This example demonstrates that the pipeline assigns the supported label not only to broad supplement-deficiency claims but also to specific dosed-drug claims when the retrieved RCT and systematic-review evidence directly substantiates them.

For **exaggerated** claims, a YouTube post titled “TURMERIC STARVES CANCER CELLS NEWER STUDY FINDS” was classified with 0.95 confidence and a critical risk tier—a marked increase in risk over the Reddit MSC example above. Five PubMed articles, three openFDA records, and an NIH fact sheet were retrieved; while curcumin has demonstrated anti-cancer activity in laboratory and preclinical contexts, the retrieved articles documented turmeric's role in managing radiation- and chemotherapy-related complications such as radiodermatitis and oral mucositis, not in starving cancer cells in human cancer treatment. The escalation to a critical risk tier reflects that an exaggerated cancer-cure claim can plausibly motivate patients to delay or substitute for evidence-based oncologic care.

For **unsupported** claims, a Reddit post stated “Im taking it for migraine headaches” in reference to magnesium L-threonate (confidence 0.95, moderate risk). Zero PubMed articles, three openFDA records, and a generic NIH magnesium reference were retrieved; no randomized controlled trial or systematic review evaluated magnesium L-threonate specifically for migraine. The claim is thus distinguished from the kratom example above on two axes: it is on Reddit rather than YouTube, and it carries a moderate rather than low risk tier because, unlike an environmental claim, an ineffective migraine treatment can plausibly delay appropriate care.

For **contradicted** claims, a YouTube post asserted “Gabapentin and tramadol give me prostate cancer all know it did” (confidence 0.98, critical risk). Five PubMed articles and three openFDA records were retrieved; multiple high-quality randomized controlled trials documented tramadol's therapeutic use during and after prostate procedures, including biopsies, with no oncogenic effect reported in any retrieved study. This example complements the ivermectin case by showing that the contradicted label applies symmetrically: it captures both false claims of efficacy (ivermectin/COVID) and false claims of harm (tramadol/prostate cancer).

For **dangerous** claims, a YouTube narrative described a fatal outcome from intentional ingestion of 35% food-grade hydrogen peroxide (confidence 0.99, critical risk). The verbatim claim was: “i was notified of the death of my good friend. I asked what had happened and was shocked to learn she had purposely ingested Hydrogen peroxide!”. Zero PubMed articles, zero openFDA records, and no NIH reference were retrieved for this specific use, but the model flagged the claim as dangerous on the basis of well-established toxicology: 35% hydrogen peroxide is corrosive and causes systemic gas embolism, gastrointestinal hemorrhage, and respiratory distress when ingested, with documented fatalities. Together with the livestock dewormer example, this illustrates that the dangerous label is triggered both by veterinary-only off-label use and by ingestion of a substance with established human toxicity, on both Reddit and YouTube.

Across the ten examples, the pipeline assigns labels and risk tiers consistently when the same product class appears on different platforms, and the risk tier scales with the plausibility and severity of the harm rather than with surface text features. The full classification records, including the complete reasoning text and supporting evidence for every claim in the dataset, are publicly available in the project repository.

5. Discussion

This study developed and applied a retrieval-augmented generation (RAG) pipeline for detecting and classifying health product misinformation on social media, combining LLM-based claim extraction with evidence grounding from PubMed, openFDA adverse event data, and NIH/NCCIH fact sheets. The results demonstrate that the proposed approach substantially outperforms a keyword-based baseline in discriminating between different types and severities of health misinformation.

5.1. Effectiveness of Evidence-Grounded Classification

The central finding of this study is that grounding LLM classifications in retrieved biomedical evidence substantially improves discriminative capacity compared to surface-level keyword matching. The keyword baseline classified 96.8% of all claims as unsupported, effectively reducing the task to a near-trivial single-class assignment. In contrast, the LLM+RAG pipeline redistributed a substantial portion of these claims into more informative categories, identifying 14.1 times more exaggerated claims and 5.4 times more dangerous claims than the baseline. This redistribution is directly attributable to the evidence retrieval stage: among claims with at least one evidence source, only 53.5% were classified as unsupported, compared to 84.7% among claims with no evidence. The presence of evidence enabled the classifier to distinguish between claims that are partially supported but overstated (exaggerated), claims that directly contradict the literature (contradicted), and claims that pose safety risks (dangerous), distinctions that are not accessible to keyword-based methods.

The classification confidence scores further support the reliability of the evidence-grounded approach. The model expressed the highest confidence for dangerous (mean 0.94) and contradicted (mean 0.90) classifications, and the lowest for supported claims (mean 0.81). This asymmetry is consistent with the nature of the task: refuting a claim or identifying a safety risk typically involves clear negative evidence, whereas confirming a claim requires evaluating the strength and scope of supporting studies, which introduces more uncertainty.

5.2. Platform Differences in Misinformation Profiles

A notable finding is the substantial divergence in misinformation profiles between Reddit and YouTube. YouTube exhibited markedly higher rates of exaggerated

(27.0% vs. 12.4%) and dangerous (11.3% vs. 2.9%) claims, while Reddit was dominated by unsupported claims (74.9% vs. 53.0%). The proportion of claims assigned high or critical risk was nearly four times higher on YouTube (17.7%) than on Reddit (4.4%). We note that, because all veracity and risk labels were produced by the LLM+RAG pipeline rather than by independent human annotation, these cross-platform differences should be interpreted as patterns produced by a single classification model, not as ground-truth differences in the underlying content. The interpretation in the following paragraphs is consistent with these patterns, but their magnitude could shift under expert annotation.

These differences likely reflect structural and community-level factors. YouTube health content in our dataset was disproportionately sourced from practitioner-style channels that make definitive claims about supplements and alternative treatments, consistent with the higher proportion of definitive claim strength on YouTube (19.3% vs. 14.5%). Reddit health subreddits, by contrast, tend to feature more personal anecdotal reports (43.4% anecdotal vs. 37.9%), which are less likely to be classified as exaggerated or dangerous because they describe individual experiences rather than making generalizable health assertions. YouTube claims were also more concentrated in the disease cure (10.8% vs. 2.4%) and drug misuse (10.8% vs. 5.4%) categories, both of which carry inherently higher risk.

At the subreddit level, condition-specific communities such as r/PCOS, r/Fibromyalgia, and r/SkincareAddiction exhibited the highest rates of supported claims (15–21%), suggesting that users in these communities more frequently discuss treatments with established evidence. In contrast, r/AlternativeHealth and r/NaturalRemedies had the highest dangerous classification rates among Reddit subreddits (9% and 7%, respectively), consistent with the promotion of unproven remedies in these communities.

5.3. Claim Categories and Risk Patterns

The category-level analysis revealed distinct risk profiles. Disease cure claims were the most concerning category, with 30.1% classified as dangerous and 38.7% as exaggerated, while only 1.3% were supported by evidence. This pattern reflects the prevalence of definitive cure claims for serious conditions such as cancer (202 claims), COVID-19 (224), and diabetes (69), where the gap between social media assertions and scientific evidence is particularly wide. Drug misuse claims showed a similar profile, with 18.3% dangerous and 31.4% exaggerated.

Detox claims presented a distinct pattern: 86.2% unsupported with zero supported claims and 69.4% returning no evidence from any of the three databases. This finding highlights a structural limitation of the pipeline rather than a classification failure: detoxification products are largely absent from the biomedical literature, and therefore the system correctly identifies these claims as lacking evidential basis but cannot distinguish between genuinely unsupported claims and claims for which the retrieval sources simply have no coverage.

The most frequently implicated products, including ashwagandha (417 claims), kratom (402), black seed oil (371), and ivermectin (242), represent substances that span the spectrum from relatively well-studied supplements to products with established safety concerns. Kratom in particular was associated with 114 dangerous classifications (123 including variant names), the highest of any single product, consistent with its documented risks including dependence, liver toxicity, and adverse drug interactions.

5.4. Role of Evidence Sources

The three evidence sources contributed complementary information. PubMed provided the highest-quality evidence but covered only 42.9% of claims, reflecting the reality that many products discussed on social media have limited peer-reviewed literature.

The openFDA adverse event database had the broadest coverage (67.9%), providing safety signals even for products without formal clinical trials. NIH and NCCIH fact sheets, while covering only 38.9% of claims, provided curated authoritative assessments for common supplements. The combination of all three sources achieved 79.4% overall evidence coverage, with 1513 claims matched by all three simultaneously.

The evidence retrieval was biased toward high-quality study designs by filtering PubMed searches for systematic reviews, randomized controlled trials, and meta-analyses. This design choice prioritizes the reliability of evidence at the cost of recall: some claims about less-studied products may lack PubMed coverage not because the claim is false, but because the relevant research has not been conducted at a level that meets the filter criteria.

5.5. Limitations

Several limitations should be considered when interpreting these results. First, the pipeline relies on LLM-generated classifications without external validation against expert annotations. While the classification confidence scores are generally high (mean 0.87) and the results are internally consistent, the absence of a human-annotated ground truth means that the reported veracity distributions reflect the model's judgments rather than independently verified labels. Future work should incorporate expert annotation of a stratified random sample to estimate classification accuracy, inter-rater agreement, and systematic biases.

Second, the veracity classification is bounded by the coverage and scope of the evidence retrieval sources. As reported in Section 4.4, the unsupported category itself is heterogeneous: 49.3% of unsupported claims (2438 of 4945) returned no evidence from any database, while 50.7% (2507 of 4945) returned at least one record that did not substantiate the claim. The first group reflects a coverage limitation of the retrieval sources rather than evidence against the claim, and should not be interpreted as evidence of falsehood; the second group reflects substantive disagreement between the asserted effect and the available literature. The detox category exemplifies the first situation: 69.4% of detox claims returned no evidence, and 86.2% were classified as unsupported, likely because detoxification products are largely absent from the biomedical literature rather than because they have been demonstrably refuted. Expanding the evidence sources to include clinical trial registries, Cochrane reviews, and pharmacovigilance databases could improve coverage and reduce the size of the no-evidence sub-category in particular.

Third, the dataset is limited to two platforms, Reddit and YouTube, and to English-language content. Health misinformation patterns may differ substantially on other platforms, such as TikTok, Instagram, Facebook, or X, and in non-English-language communities.

Fourth, the claim extraction step relies entirely on a single LLM (Claude Haiku) without a second-pass verification or human review. The 8250 extracted claims include three instances of an inconsistent label ("anecdotic" instead of "anecdotal"), and the extraction may miss implicit claims or misinterpret sarcasm and irony, which are common in informal social media text.

Fifth, although we evaluated the LLM+RAG pipeline against three complementary baselines—a keyword classifier, a zero-shot pretrained transformer without retrieved evidence, and the same transformer with retrieved evidence—all three are zero-shot or rule-based, and are designed primarily to isolate the contributions of evidence retrieval and language-model capability rather than to position the pipeline against the state of the art. A more comprehensive evaluation would additionally include transformer classifiers fine-tuned on health-misinformation datasets (e.g., DistilBERT or BioBERT fine-tuned on COVID-Misinfo or similar corpora) and end-to-end misinformation detection systems,

evaluated on a shared, expert-annotated test set. We identify this comparison as a high-priority direction for future work.

Finally, the data collection captured a snapshot of content available at the time of collection (April 2026) and is subject to survivorship bias: deleted posts, removed comments, and content moderation actions may have already filtered some of the most egregious misinformation from the dataset.

6. Conclusions

This study presented a retrieval-augmented generation pipeline for detecting and classifying health product misinformation on social media. By grounding LLM-based veracity classification in evidence from PubMed, openFDA adverse event reports, and NIH/NCCIH fact sheets, the pipeline assigned a substantially more fine-grained distribution of veracity labels than three evidence-free or evidence-only baselines: a keyword classifier that defaulted 96.8% of claims to UNSUPPORTED; a zero-shot transformer (facebook/bart-large-mnli [45]) that without evidence collapsed 85.1% of claims to EXAGGERATED (accuracy 21.7%, macro F1 0.131); and the same transformer with retrieved evidence which doubled accuracy to 47.6% (macro F1 0.225) but still failed to detect EXAGGERATED, DANGEROUS, or CONTRADICTED claims (per-class F1 below 0.10 for each). The three-baseline ablation indicates that evidence retrieval and a language model capable of reasoning over the retrieved evidence are jointly necessary for fine-grained misinformation classification; neither component is sufficient on its own. We emphasize that these counts are model-assigned rather than independently validated, and they should be interpreted as illustrating the discriminative behavior of an evidence-grounded LLM relative to evidence-free and smaller-LLM baselines, pending external validation against human or expert annotation. Analysis of 8250 health claims extracted from 3493 Reddit and YouTube posts revealed that only 7.0% of claims were supported by biomedical evidence, while 33.1% were classified as exaggerated, contradicted, or dangerous. YouTube carried a disproportionate share of high-risk misinformation compared to Reddit, and disease cure claims were the most frequently flagged category for dangerous content. The most commonly implicated products—ashwagandha, kratom, black seed oil, turmeric, and ivermectin—represent substances that span the spectrum from relatively well-studied supplements to products with documented safety concerns, and continue to be promoted with unsubstantiated health claims on social media platforms. These findings underscore the potential of evidence-grounded LLM pipelines as tools for health misinformation surveillance, while also highlighting the need for expanded evidence coverage, expert validation, and cross-platform generalization in future work.

Author Contributions: Methodology, S.B. and Z.S. (Zeinab Shahbazi); validation, Z.S. (Zeinab Shahbazi), S.B., S.J. and Z.S. (Zahra Shahbazi); formal analysis, Z.S. (Zeinab Shahbazi) and S.J.; resources, S.J. and Z.S. (Zeinab Shahbazi); data curation, Z.S. (Zeinab Shahbazi), S.B. and S.J.; writing—original draft preparation, Z.S. (Zeinab Shahbazi), S.B., Z.S. (Zahra Shahbazi) and S.J.; writing—review and editing, Z.S. (Zeinab Shahbazi) and S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The pipeline source code is available at <https://github.com/sarabehnamian/health-misinfo-detection>, accessed on 10 May 2026. Raw social media data are not included due to platform terms of service.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Suarez-Lledo, V.; Alvarez-Galvez, J. Prevalence of Health Misinformation on Social Media: Systematic Review. *J. Med. Internet Res.* **2021**, *23*, e17187. [\[CrossRef\]](#)
2. Malki, L.M.; Patel, D.; Singh, A. “The Headline Was So Wild That I Had To Check”: An Exploration of Women’s Encounters with Health Misinformation on Social Media. *Proc. ACM Hum.-Comput. Interact.* **2024**, *8*, 1–26. [\[CrossRef\]](#)
3. Kisa, S.; Kisa, A. A Comprehensive Analysis of COVID-19 Misinformation, Public Health Impacts, and Communication Strategies: Scoping Review. *J. Med. Internet Res.* **2024**, *26*, e56931. [\[CrossRef\]](#)
4. Do Nascimento, I.J.B.; Pizarro, A.B.; Almeida, J.M.; Azzopardi-Muscat, N.; Gonçalves, M.A.; Björklund, M.; Novillo-Ortiz, D. Infodemics and Health Misinformation: A Systematic Review of Reviews. *Bull. World Health Organ.* **2022**, *100*, 544. [\[CrossRef\]](#)
5. Clark, O.; Joshi, K.P. Evaluating Causal AI Techniques for Health Misinformation Detection. In *Proceedings of the 2025 IEEE International Conference on Pervasive Computing and Communications Workshops and Other Affiliated Events (PerCom Workshops), Washington, DC, USA, 17–21 March 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 50–55.
6. Alkhalwaldeh, F. False Textual Information Detection, a Deep Learning Approach. Ph.D. Thesis, University of York, York, UK, 2022.
7. Allam, H.; Makubvure, L.; Gyamfi, B.; Graham, K.N.; Akinwolere, K. Text Classification: How Machine Learning Is Revolutionizing Text Categorization. *Information* **2025**, *16*, 130. [\[CrossRef\]](#)
8. Mohasseb, A.; Amer, E.; Chiroma, F.; Tranchese, A. Leveraging Advanced NLP Techniques and Data Augmentation to Enhance Online Misogyny Detection. *Appl. Sci.* **2025**, *15*, 856. [\[CrossRef\]](#)
9. Del Vicario, M.; Bessi, A.; Zollo, F.; Petroni, F.; Scala, A.; Caldarelli, G.; Stanley, H.E.; Quattrociocchi, W. The Spreading of Misinformation Online. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 554–559. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Cole, J.; Watkins, C.; Kleine, D. Health Advice from Internet Discussion Forums: How Bad Is Dangerous? *J. Med. Internet Res.* **2016**, *18*, e4. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Swire-Thompson, B.; Lazer, D. Public Health and Online Misinformation: Challenges and Recommendations. *Annu. Rev. Public Health* **2020**, *41*, 433–451. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Beres, D.; Remski, M.; Walker, J. *Conspiracy: How New Age Conspiracy Theories Became a Health Threat*; PublicAffairs: New York, NY, USA, 2023.
13. Xie, Q.; Schenck, E.J.; Yang, H.S.; Chen, Y.; Peng, Y.; Wang, F. Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond. *medRxiv* **2023**, 2023.04.18.23288752. [\[CrossRef\]](#)
14. Kou, Z.; Shang, L.; Zhang, Y.; Wang, D. HC-COVID: A Hierarchical Crowdsourced Knowledge Graph Approach to Explainable COVID-19 Misinformation Detection. *Proc. ACM Hum.-Comput. Interact.* **2022**, *6*, 1–25. [\[CrossRef\]](#)
15. Cui, L.; Seo, H.; Tabar, M.; Ma, F.; Wang, S.; Lee, D. DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation. In *KDD ’20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 492–502.
16. Hu, X.; Chen, J.; Li, X.; Guo, Y.; Wen, L.; Yu, P.S.; Guo, Z. Do Large Language Models Know about Facts? *arXiv* **2023**, arXiv:2310.05177. [\[CrossRef\]](#)
17. Zhang, T.; Luo, H.; Chuang, Y.S.; Fang, W.; Gaitskell, L.; Hartvigsen, T.; Wu, X.; Fox, D.; Meng, H.; Glass, J. Interpretable Unified Language Checking. *arXiv* **2023**, arXiv:2304.03728. [\[CrossRef\]](#)
18. Van Dis, E.A.; Bollen, J.; Zuidema, W.; Van Rooij, R.; Bockting, C.L. ChatGPT: Five Priorities for Research. *Nature* **2023**, *614*, 224–226. [\[CrossRef\]](#)
19. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv* **2023**, arXiv:2303.13375. [\[CrossRef\]](#)
20. Su, Q.; Wan, M.; Liu, X.; Huang, C.R. Motivations, Methods and Metrics of Misinformation Detection: An NLP Perspective. *Nat. Lang. Process. Res.* **2020**, *1*, 1–13. [\[CrossRef\]](#)
21. Mallik, A.; Kumar, S. Word2Vec and LSTM-Based Deep Learning Technique for Context-Free Fake News Detection. *Multimed. Tools Appl.* **2024**, *83*, 919–940. [\[CrossRef\]](#)
22. Al-Tarawneh, M.A.; Al-ir, O.; Al-Maaitah, K.S.; Kanj, H.; Aly, W.H.F. Enhancing Fake News Detection with Word Embedding: A Machine Learning and Deep Learning Approach. *Computers* **2024**, *13*, 239. [\[CrossRef\]](#)
23. Wang, J.; Wang, X.; Yu, A. Tackling Misinformation in Mobile Social Networks: A BERT-LSTM Approach for Enhancing Digital Literacy. *Sci. Rep.* **2025**, *15*, 1118. [\[CrossRef\]](#)
24. Kukreja, S.; Kumar, T.; Purohit, A.; Dasgupta, A.; Guha, D. A Literature Survey on Open Source Large Language Models. In *ICCMB ’24: Proceedings of the 2024 7th International Conference on Computers in Management and Business*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 133–143.

25. Bai, G.; Chai, Z.; Ling, C.; Wang, S.; Lu, J.; Zhang, N.; Shi, T.; Yu, Z.; Zhu, M.; Zhang, Y.; et al. Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models. *arXiv* **2024**, arXiv:2401.00625. [\[CrossRef\]](#)
26. Abu-Salih, B. Domain-Specific Knowledge Graphs: A Survey. *J. Netw. Comput. Appl.* **2021**, *185*, 103076. [\[CrossRef\]](#)
27. Liu, J.; Zhang, J.; Jan, M.A.; Sun, R.; Liu, L.; Verma, S.; Chatterjee, P. A Comprehensive Privacy-Preserving Federated Learning Scheme with Secure Authentication and Aggregation for Internet of Medical Things. *IEEE J. Biomed. Health Inform.* **2023**, *28*, 3282–3292. [\[CrossRef\]](#)
28. Meduri, K.; Nadella, G.S.; Yadulla, A.R.; Kasula, V.K.; Maturi, M.H.; Brown, S.; Satish, S.; Gonaygunta, H. Leveraging Federated Learning for Privacy-Preserving Analysis of Multi-Institutional Electronic Health Records in Rare Disease Research. *J. Econ. Technol.* **2025**, *3*, 177–189. [\[CrossRef\]](#)
29. Abaoud, M.; Almuqrin, M.A.; Khan, M.F. Advancing Federated Learning through Novel Mechanism for Privacy Preservation in Healthcare Applications. *IEEE Access* **2023**, *11*, 83562–83579. [\[CrossRef\]](#)
30. Zhou, I.; Tofigh, F.; Piccardi, M.; Abolhasan, M.; Franklin, D.; Lipman, J. Secure Multi-Party Computation for Machine Learning: A Survey. *IEEE Access* **2024**, *12*, 53881–53899. [\[CrossRef\]](#)
31. Jonnagaddala, J.; Wong, Z.S.Y. Privacy-Preserving Strategies for Electronic Health Records in the Era of Large Language Models. *npj Digit. Med.* **2025**, *8*, 34. [\[CrossRef\]](#)
32. Papageorgiou, E.; Chronis, C.; Varlamis, I.; Himeur, Y. A Survey on the Use of Large Language Models (LLMs) in Fake News. *Future Internet* **2024**, *16*, 298. [\[CrossRef\]](#)
33. Lucas, J.; Uchendu, A.; Yamashita, M.; Lee, J.; Rohatgi, S.; Lee, D. Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023*; pp. 14279–14305.
34. Chen, C.; Shu, K. Combating Misinformation in the Age of LLMs: Opportunities and Challenges. *AI Mag.* **2024**, *45*, 354–368. [\[CrossRef\]](#)
35. Shahbazi, Z.; Behnamian, S. Using Large Language Models to Detect and Debunk Climate Change Misinformation. *Big Data Cogn. Comput.* **2026**, *10*, 34. [\[CrossRef\]](#)
36. Pendyala, V.S.; Hall, C.E. Explaining Misinformation Detection Using Large Language Models. *Electronics* **2024**, *13*, 1673. [\[CrossRef\]](#)
37. Kumari, S.; Reddy, H.K.; Kulkarni, C.S.; Gowthami, V. Debunking Health Fake News with Domain-Specific Pre-Trained Model. *Glob. Transit. Proc.* **2021**, *2*, 267–272. [\[CrossRef\]](#)
38. Ayoub, J.; Yang, X.J.; Zhou, F. Combat COVID-19 Infodemic Using Explainable Natural Language Processing Models. *Inf. Process. Manag.* **2021**, *58*, 102569. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Alghamdi, J.; Lin, Y.; Luo, S. Towards COVID-19 Fake News Detection Using Transformer-Based Models. *Knowl.-Based Syst.* **2023**, *274*, 110642. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Upadhyay, R.; Pasi, G.; Viviani, M. Leveraging Socio-Contextual Information in BERT for Fake Health News Detection in Social Media. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks (OASIS '23), Rome, Italy, 4–8 September 2023*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 38–46. [\[CrossRef\]](#)
41. Men, X.; Mariano, V.Y. Explainable Fake News Detection Based on BERT and SHAP Applied to COVID-19. *Int. J. Mod. Educ. Comput. Sci.* **2024**, *16*, 11–22. [\[CrossRef\]](#)
42. Khemani, B.; Patil, S.; Kotecha, K.; Vora, D. Detecting Health Misinformation: A Comparative Analysis of Machine Learning and Graph Convolutional Networks in Classification Tasks. *MethodsX* **2024**, *12*, 102737. [\[CrossRef\]](#)
43. Padalko, H.; Chomko, V.; Yakovlev, S.; Chumachenko, D. A Novel Comprehensive Framework for Detecting and Understanding Health-Related Misinformation. *Information* **2025**, *16*, 175. [\[CrossRef\]](#)
44. Gargari, O.K.; Habibi, G. Enhancing Medical AI with Retrieval-Augmented Generation: A Mini Narrative Review. *Digit. Health* **2025**, *11*, 20552076251337177. [\[CrossRef\]](#)
45. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 7871–7880.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.