

Health Misinformation Detection in Hidradenitis Suppurativa Communities Using Multi-Model LLM Ensembles

Jayasimha Shivanna¹, Shagun Saboo², Dhraastiben Ankurbhai Bhavasar¹, Divya Chaudhary¹

¹ Khoury College of Computer Sciences, Northeastern University, Seattle and Boston, USA

² Independent Researcher, Boston, USA
{shivanna.j, d.chaudhary}@northeastern.edu

Abstract

Health misinformation in online patient communities poses significant risks for individuals managing chronic conditions, yet no detection frameworks exist for rare dermatological diseases. We present the first unsupervised multi-model framework for detecting health misinformation in Hidradenitis Suppurativa (HS) Reddit communities, requiring zero human annotation. Our framework deploys 28 models across five experimental stages: aspect-based sentiment and emotion analysis using RoBERTa models, weakly supervised classification via 10 auto-labeled RoBERTa-family architectures, unsupervised pattern discovery with UMAP/HDBSCAN clustering, zero-shot detection via 5 NLI classifiers, and prompt-based detection via 5 open-source LLMs including Llama-3, Mistral-7B, Flan-T5, BioMistral, and GatorTron. We introduce the HS Misinformation Index (HSMI), a composite risk metric fusing multi-model consensus, domain-specific keyword heuristics, and contextual sentiment signals. Applied to 9,838 Reddit texts, RoBERTa-large achieves the highest supervised F1 (0.774), while Llama-3.2-3B leads among LLMs (F1 = 0.676). NLI classifiers flag 2.0% of texts by majority consensus versus 12.0% for LLMs, a gap we attribute to fundamentally different classification mechanisms: entailment-based logical judgment versus prompt-based pragmatic assessment. Inter-model agreement remains slight for both open-source paradigms ($\kappa = 0.20$ for NLI, 0.16 for LLMs), while a closed-source validation using GPT-4o, Claude Sonnet 4, and Gemini models on a seven-category taxonomy achieves substantially higher agreement ($\kappa = 0.62$). Emotion-risk analysis reveals that texts expressing disgust and anger carry the highest HSMI scores, while joyful texts carry the lowest, and general-purpose models consistently outperform domain-specific ones across all stages. This work presents the first large-scale systematic analysis of HS misinformation and demonstrates that multi-model consensus can surface meaningful misinformation patterns without human labels, providing a scalable methodology for underserved chronic disease communities. We release the dataset with model-derived annotations to support future expert-validated benchmarking.

Introduction

Hidradenitis Suppurativa (HS) is a chronic, inflammatory, and often debilitating dermatological condition characterized by recurrent painful nodules, abscesses, and sinus tract

formation, primarily affecting intertriginous areas of the body (Jemec 2012; Alikhan et al. 2009). Beyond its physical manifestations, HS imposes a profound psychosocial burden on patients, including chronic pain, reduced mobility, social isolation, stigmatization, and increased prevalence of depression and anxiety (Matusiak 2010; Kouris et al. 2016). The unpredictable disease trajectory and limited universally effective treatments further exacerbate patient distress, often leaving individuals navigating their condition with uncertainty and frustration (Saunte and Jemec 2017). Consequently, HS is not merely a dermatological disorder but a life-altering condition that deeply impacts patients' quality of life and mental well-being.

In recent years, social media platforms have transformed the healthcare information ecosystem, enabling patients to share experiences, seek advice, and build support communities (Moorhead et al. 2013; De Choudhury 2014). Among these, Reddit stands out due to its pseudonymous nature, allowing individuals to discuss sensitive health conditions without fear of stigma, leading to more candid and emotionally expressive conversations (De Choudhury et al. 2015). For individuals suffering from HS, these platforms have become critical spaces to document symptoms, discuss treatment journeys, and voice the persistent pain and challenges they endure. While such discussions offer valuable insight into patient experiences, they also expose a parallel concern the widespread sharing of unverified and potentially harmful medical advice (Fox 2013). Users frequently recommend home remedies, off-label treatments, or lifestyle changes without clinical validation, which may delay appropriate medical care or encourage unsafe self-treatment. Despite growing reliance on social media for health information, systematic analysis of HS-related discourse remains limited. Prior work has explored similar dynamics in other chronic and autoimmune conditions, highlighting both support and misinformation (Perez-Chada 2019; Guidry 2017; Wang et al. 2019b), and broader studies have emphasized misinformation as a critical public health issue (Zarocostas 2020). However, there is still a lack of structured frameworks to analyze and quantify misinformation within HS-specific patient communities, leaving a significant gap in understanding its impact on patient decision-making.

The motivation for this work stems from the urgent need to address this gap. Accurate classification and analysis

of information within HS-related social media discussions can enable early identification of harmful advice, promote safer online communities, and protect patients from dangerous self-treatment practices. Moreover, such insights can inform evidence-based strategies for patient education, improve clinician awareness of patient concerns, and enhance counseling approaches by aligning them with real-world patient narratives. By bridging the divide between clinical knowledge and patient discourse, this research aims to contribute toward a more informed and supportive digital healthcare ecosystem.

To achieve this, we collect and analyze a large corpus of HS-related posts from Reddit and employ a comprehensive suite of both open-source and closed-source machine learning models to classify and interpret the content. Our methodology incorporates zero-shot natural language inference (NLI) classifiers, sentiment and emotion detection models, embedding-based semantic representations, and large language models (LLMs). Specifically, we utilize models such as BART-MNLI, RoBERTa-MNLI, DeBERTa-v3 variants, and XLM-RoBERTa for zero-shot classification; sentiment models including Twitter-RoBERTa and DistilBERT-SST2; emotion classifiers such as DistilRoBERTa-Emotion and GoEmotions; and embedding models like All-RoBERTa-Large-v1. Additionally, we leverage open-source LLMs including Llama-3-8B-Instruct, Mistral-7B-Instruct, Flan-T5-Large, BioMistral-7B, and GatorTron-Base, alongside fine-tuned RoBERTa-family architectures for domain adaptation. For comparative analysis, we further incorporate state-of-the-art closed-source models such as GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, and Med-PaLM 2 to evaluate their effectiveness in identifying and classifying misinformation within patient-generated content.

The remainder of this paper is structured as follows: we review relevant literature, detail our methodology and modeling approaches, evaluate the comparative performance of open-source and closed-source models in misinformation detection, and conclude with a discussion of findings, clinical implications, and future research directions.

Related Work

Hidradenitis Suppurativa (HS) has been extensively studied from clinical, epidemiological, and pathophysiological perspectives. Foundational work has characterized HS as a chronic inflammatory disease with complex etiology involving genetic susceptibility, immune dysregulation, and environmental triggers (Alikhan et al. 2009; Sabat et al. 2024; Frew 2024). Large-scale studies further highlight its global burden and associated comorbidities, including metabolic syndrome, cardiovascular risks, and autoimmune conditions (Hay et al. 2014; Garg et al. 2017; Nielsen et al. 2024). Recent meta-analyses estimate HS prevalence and emphasize its underdiagnosis across populations (Bouazzi et al. 2024). Importantly, HS has been consistently linked to severe psychosocial consequences, including depression, anxiety, and reduced quality of life (Dalgard et al. 2015; Szepietowska et al. 2026). While these studies provide critical clinical insights, they primarily focus on biological and epidemiological dimensions, leaving patient-driven narratives partic-

ularly those emerging from digital platforms largely unexplored. Our work builds on this clinical understanding by shifting focus toward how patients articulate their experiences in online communities and how such discourse may influence health behaviors.

The role of social media in healthcare has grown significantly, with platforms enabling patients to share experiences, seek support, and access information beyond traditional clinical settings. Early work demonstrated how online communities contribute to patient empowerment and peer support (Moorhead et al. 2013; Fox 2013), while later studies highlighted their utility in monitoring public health trends and mental health signals (Sarker et al. 2020; De Choudhury 2014). However, the same platforms have also become conduits for misinformation, with studies documenting the rapid spread of false or misleading health content (Chou et al. 2018; Wang et al. 2019a). Large-scale analyses of information diffusion reveal that false information spreads faster and more widely than factual content (Vosoughi, Roy, and Aral 2018; Cinelli et al. 2020). Research during the COVID-19 pandemic further quantified the prevalence and impact of misinformation, emphasizing its risks to public health decision-making (Kouzy et al. 2020; Patwa et al. 2021). While these studies establish the dual nature of social media as both a support system and a misinformation vector, they largely focus on broad public health contexts rather than condition-specific communities such as HS. Our work extends this line of research by focusing on a niche yet highly vulnerable patient group and analyzing the nature of information exchanged within it.

Several studies have explored the intersection of social media and chronic or autoimmune diseases, demonstrating how patients use online platforms to share experiences, discuss treatments, and navigate uncertainties. For instance, research on psoriasis, rheumatoid arthritis, and other autoimmune conditions shows that social media serves as a critical space for peer-to-peer knowledge exchange but also introduces risks of unverified treatment recommendations (Perez-Chada 2019; Wang et al. 2019b; Guidry 2017). Investigations into Reddit-based health communities reveal rich patient narratives, including symptom descriptions, treatment journeys, and emotional expressions, often absent from clinical records (De Choudhury et al. 2015). However, these studies primarily focus on understanding patient sentiment or community dynamics rather than systematically evaluating the accuracy of shared information. In the context of HS, while anecdotal evidence suggests a strong presence of patient communities on platforms like Reddit, there is a lack of structured research analyzing the content, themes, and potential risks within these discussions. Our work addresses this gap by systematically studying HS-related discourse and explicitly focusing on misinformation as a core research problem.

From a methodological standpoint, natural language processing (NLP) techniques have been widely used to analyze social media data, including topic modeling, sentiment analysis, and misinformation detection. Traditional approaches such as Latent Dirichlet Allocation (LDA) have been employed to uncover thematic structures in large text

corpora (Blei et al. 2003), while lexicon-based and early machine learning methods like VADER have enabled sentiment analysis in social media contexts (Hutto and Gilbert 2014). The advent of transformer-based models, particularly BERT and its variants, has significantly advanced text understanding capabilities (Devlin et al. 2019). Domain-specific adaptations such as BioBERT and BioGPT further enhance performance in biomedical contexts (Lee et al. 2020; Luo et al. 2022). Additionally, emotion detection frameworks and shared tasks have enabled fine-grained analysis of affective signals in text (Mohammad et al. 2018; Acheampong et al. 2020). While these models have been applied to various healthcare datasets, their use in analyzing HS-specific social media content remains unexplored. In this work, we leverage these advancements to perform multi-dimensional analysis of HS discourse, including classification, sentiment, and emotional context.

Misinformation detection has emerged as a critical research area, with studies proposing various models and frameworks to identify false or misleading content. Early work focused on feature-based and probabilistic approaches, while more recent methods employ deep learning architectures and hybrid models (Shu et al. 2017; Zhou and Zafarani 2020). Techniques such as CSI and other neural approaches incorporate content, user behavior, and propagation patterns to detect fake news (Ruchansky et al. 2017). Benchmark datasets and shared tasks, particularly during the COVID-19 infodemic, have accelerated progress in this domain (Patwa et al. 2021). However, most misinformation research is centered around news articles or large-scale public health narratives rather than patient-generated content in niche communities. Furthermore, the application of these techniques to domain-specific discussions, such as HS, where misinformation may be subtle, anecdotal, or experience-driven, remains limited. Our work adapts and extends these methodologies to the context of HS, focusing on identifying misinformation embedded within patient narratives rather than overtly false claims.

Recent advancements in large language models (LLMs) have further transformed text analysis and reasoning capabilities. Models such as GPT-4 (OpenAI 2023), Med-PaLM (Singhal et al. 2023), and other transformer-based architectures demonstrate strong performance in clinical reasoning and language understanding. Research on zero-shot and few-shot learning highlights the ability of these models to generalize across tasks without extensive fine-tuning (Brown et al. 2020; Kojima et al. 2022). Frameworks such as Hugging-GPT illustrate the integration of multiple models for complex task pipelines (Shen et al. 2023). Additionally, studies on self-consistency and reasoning strategies further enhance model reliability (Wang et al. 2023). Despite these advancements, there remains limited work comparing open-source and closed-source models in the context of healthcare misinformation detection, particularly within social media data. Our work addresses this gap by conducting a comprehensive comparative analysis of both open-source and proprietary models including GPT-4o, Claude, Gemini, and Med-PaLM on HS-related Reddit data. In summary, prior research has extensively explored HS from a clinical perspective, so-

cial media as a healthcare tool, and misinformation detection using advanced NLP techniques. However, there is a notable absence of studies that integrate these domains to analyze misinformation within HS-specific social media discussions. To the best of our knowledge, no prior work has systematically examined the nature, prevalence, and impact of misinformation in HS communities on platforms like Reddit using a combination of modern NLP and LLM-based approaches. By addressing this gap, our research aims to provide actionable insights into patient discourse, contribute to safer online health communities, and advance the application of AI in domain-specific healthcare analysis.

Methodology

The approach we take in this work is shaped by a basic constraint: there are no labeled datasets for HS-related misinformation, and creating one through expert annotation would require clinical knowledge that is difficult to scale. Rather than treating this as a limitation, we designed a framework that works without labels entirely. The core idea is straightforward. If multiple independent models, each trained on different data and using different architectures, agree that a particular post looks like misinformation, that agreement itself becomes a meaningful signal. No single model needs to be correct on its own. What matters is the pattern of convergence.

The framework is organized into five stages, each building on the previous one. We begin with data collection and preprocessing, move through aspect-level discourse analysis, then explore both supervised and unsupervised detection strategies, and finally compare two fundamentally different paradigms for flagging misinformation: zero-shot natural language inference and prompt-based large language models. Throughout, we use a shared set of agreement metrics so that results across stages can be compared directly as depicted in Figure 1.

Data Collection and Preprocessing

We collected 9,838 texts from the r/Hidradenitis subreddit using the Reddit API. This includes 963 original posts and 8,875 comments. The subreddit is the largest English-language online community focused on HS, and the discussions there cover a wide range of topics: symptom descriptions, treatment experiences, emotional struggles, advice-seeking, and, inevitably, claims that range from well-intentioned but inaccurate to potentially harmful.

Raw Reddit text is messy. Posts contain URLs, markdown formatting, emoji, informal contractions, bot-generated signatures, and various Unicode artifacts. To normalize this input for transformer models, we applied a seven-step cleaning pipeline. URLs and Reddit-specific artifacts were removed first. Emoji were converted to their textual descriptions rather than being discarded, since they carry emotional signal that matters for sentiment and emotion analysis. Contractions were expanded to reduce tokenization inconsistencies. Unicode characters were normalized via NFKD decomposition. Whitespace was collapsed, and texts shorter than 20 characters were filtered out as they carry too little signal for meaningful inference.

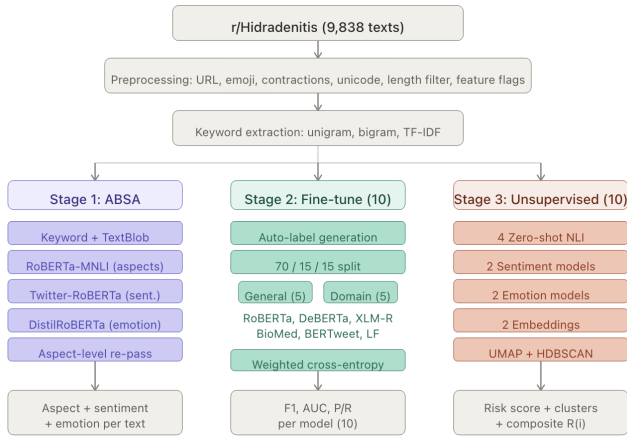


Figure 1: Data from r/Hidradenitis (9,838 texts) flows through preprocessing into three parallel stages: ABSA via 3 RoBERTa models, supervised fine-tuning of 10 models (5 general, 5 domain), and unsupervised analysis via 10 models with UMAP/HDBSCAN clustering.

Beyond cleaning, we computed three binary feature flags for each text using domain-informed regular expressions. These flags indicate whether a text mentions symptoms (terms like “flare”, “abscess”, “tunnel”), treatments (“Humira”, “Bimzelx”, “surgery”), or emotional states (“pain”, “depressed”, “anxious”). These are not used as model inputs directly but serve as lightweight indicators for corpus characterization and downstream validation.

Keyword Frequency Analysis

Before deploying any transformer models, we conducted a corpus-level keyword analysis to understand what HS patients actually talk about. After stopword removal and lemmatization, we extracted unigrams, bigrams, and trigrams, ranking them by both raw frequency and TF-IDF scores. We also organized keywords into five domain-specific categories: Symptoms, Treatments, Emotions, Lifestyle, and Medical Procedures. This step is not glamorous, but it turned out to be surprisingly useful. The patterns that emerged here directly informed the misinformation keyword heuristics we use in later stages, and they provide interpretive context for the model-based results.

Stage 1: Aspect-Based Sentiment and Emotion Analysis

HS discussions are not monolithic. A single post might express frustration about a failed treatment, gratitude toward a supportive community member, and anxiety about an upcoming surgery, all in the same paragraph. To capture this complexity, we performed aspect-based sentiment analysis (ABSA) at two levels.

The first is a lightweight keyword-matching approach paired with TextBlob for sentiment scoring. We defined two manually curated dictionaries: one mapping texts to 10 aspect categories (such as “pain and physical symp-

oms”, “treatment and medication”, “mental health and emotional wellbeing”) and another mapping to 10 emotion categories (such as “frustrated”, “hopeful”, “anxious or scared”). TextBlob provides a continuous polarity score between -1 and +1, which we map to positive, neutral, or negative labels. This approach runs in seconds on a CPU and gives us corpus-wide distributional baselines.

The second level uses three RoBERTa-family models for deeper, context-aware analysis. For aspect detection, we use roberta-large-mnli as a zero-shot classifier, evaluating each text against the 10 candidate aspect labels through natural language inference. For sentiment, we use Twitter-RoBERTa, a model fine-tuned on approximately 124 million tweets, which handles informal social media language better than models trained on formal text. For emotion, we use DistilRoBERTa-Emotion, which produces 7-class Ekman emotion predictions. We also add a fourth step that most ABSA pipelines skip: aspect-conditioned sentiment. Here, we prepend the detected aspect as context (for example, “Regarding treatment and medication: [original text]”) and re-run the sentiment model. This matters because a post that is overall neutral might express distinctly negative sentiment specifically about its treatment experience, and that distinction is lost without aspect, as seen in Figure 1.

Stage 2: Supervised Classification with Auto-Labeled RoBERTa Models

This stage is designed as a controlled baseline. In the absence of ground-truth annotations, we generate pseudo-labels by combining three complementary signals. First, we use a zero-shot classification score from BART-MNLI to estimate the likelihood that a given text contains an unverified health claim. Second, we introduce a keyword-based heuristic score, implemented through weighted regular expression matching across four categories of misinformation: cure promises, anti-medical rhetoric, pseudoscience, and commercial promotion. The weighting scheme is intentionally asymmetric: commercial promotion (weight 0.60) is penalized more heavily than cure promises (weight 0.35), reflecting the distinction between deliberate exploitation and optimistic exaggeration. Third, we incorporate a debunking detector that identifies content explicitly refuting misinformation (e.g., statements such as “there is no cure” or references to “snake oil”) to avoid incorrect labeling. The final composite score is thresholded at the 80th percentile to assign pseudo-labels.

Using these pseudo-labels, we fine tune ten models from the RoBERTa family, grouped to enable a controlled comparison. The general purpose group includes RoBERTa-base, RoBERTa-large, DistilRoBERTa, XLM-RoBERTa, and DeBERTa-v3. The domain-specific group comprises Twitter-RoBERTa (social media), BioMed-RoBERTa (biomedical literature), BERTweet (tweets), Longformer (long documents), and Clinical-Longformer (clinical notes). All models are trained using a 70/15/15 stratified train-validation-test split, with early stopping, weighted cross-entropy loss to address class imbalance, and FP16 mixed-precision training.

We emphasize that this stage does not aim to claim detection of ground-truth misinformation, as the labels are automatically generated rather than expert-annotated. Instead, the objective is to evaluate which model architectures and pretraining domains are most effective under weak supervision, thereby informing future work once high-quality annotated datasets become available.

Stage 3: Unsupervised Pattern Discovery

This is the stage where we move beyond classification and ask a different question: what patterns exist in the data that we might not have thought to look for? Ten RoBERTa-family models are deployed in inference-only mode across four analytical layers, each capturing a different type of signal.

The first layer uses four zero-shot NLI classifiers (RoBERTa-MNLI, BART-MNLI, DeBERTa-NLI, and DeBERTa-Zeroshot) to score each text against five candidate hypotheses, including “unverified health claim or misinformation.” These four models were chosen specifically for their architectural diversity: mixing encoder-only and encoder-decoder architectures, standard and disentangled attention mechanisms. If all four independently flag the same text, that convergence is hard to dismiss as a single-model artifact, as seen in Figure 2.

The second layer applies two sentiment models (Twitter-RoBERTa and Siebert-RoBERTa) with a majority-vote consensus. The third layer uses two emotion models operating at different granularity levels: DistilRoBERTa-Emotion with 7 Ekman categories and GoEmotions with 28 fine-grained categories. The fourth layer extracts sentence embeddings from two models (All-RoBERTa-Large and All-DistilRoBERTa), applies UMAP dimensionality reduction, and performs HDBSCAN density-based clustering to discover latent topic structure. Running both embedding models through identical UMAP and HDBSCAN pipelines independently provides cross-model validation: if the same high-risk texts cluster together regardless of which embedding model produced the vectors, the structure is genuine rather than model-dependent.

These signals are then fused into what we call the HS Misinformation Index (HSMI), a composite risk score defined as:

$$HSMI_i = 0.50 \cdot \bar{s}_i + 0.30 \cdot k_i + 0.20 \cdot b_i - 0.15 \cdot d_i$$

where \bar{s}_i is the mean misinformation score across the four zero-shot models, k_i is the keyword heuristic score, b_i is a positive-sentiment boost applied when optimistic framing co-occurs with misinformation-associated keywords (capturing the empirically observed tendency of false health claims to adopt hopeful language), and d_i is a debunking penalty that down-weights texts refuting misinformation rather than promoting it. The weights reflect a deliberate prioritization: the zero-shot ensemble receives the highest weight (0.50) because it is the most generalizable signal, while keyword heuristics contribute domain-specific precision (0.30) and sentiment provides contextual calibration (0.20). Risk levels are assigned using percentile-based

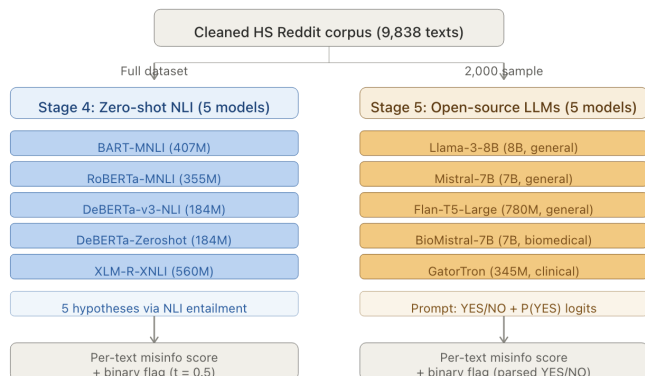


Figure 2: Stage 4 applies 5 NLI classifiers via entailment on the full dataset; Stage 5 prompts 5 open-source LLMs (7B+ models use 4-bit quantization) on a 2,000-text sample. Both produce per-text misinformation scores and binary flags.

thresholds rather than fixed cutoffs, so the framework adapts to the data distribution, as seen in Figure 4.

Stage 4: Zero-Shot NLI Misinformation Detection

While Stage 3 uses NLI models as one layer within a larger system, Stage 4 isolates the NLI paradigm to evaluate it on its own terms. Five NLI classifiers namely BART-MNLI, RoBERTa-MNLI, DeBERTa-v3-NLI, DeBERTa-Zeroshot, and XLM-R-XNLI are used to classify each of the 9,838 texts against five candidate hypotheses via textual entailment. A binary misinformation flag is set at threshold 0.5, and majority consensus requires agreement from at least three of the five models.

The choice of five models rather than three or seven is deliberate. Five provides enough diversity for meaningful consensus while keeping the agreement metrics (particularly Fleiss’ Kappa) interpretable. The models span parameter counts from 184M to 560M and include both encoder-only and encoder-decoder architectures, so agreement between them cannot be attributed to shared inductive biases.

Stage 5: Open-Source LLM Misinformation Detection

The final stage tests a fundamentally different paradigm. Instead of using pre-trained classifiers, we prompt five open-source large language models to make explicit yes-or-no judgments about whether a text contains health misinformation. The models are Llama-3-8B, Mistral-7B, Flan-T5-Large, BioMistral-7B (Mistral continued-pretrained on PubMed Central), and GatorTron (pretrained on over 90 billion words of clinical text). This selection deliberately mixes general-purpose models with domain-specific ones to test whether biomedical or clinical pretraining provides an advantage for detecting misinformation in informal patient discourse.

The practical challenges of running 7B-parameter models on a single T4 GPU are handled through 4-bit NF4 quantization via bitsandbytes. Each causal language model re-

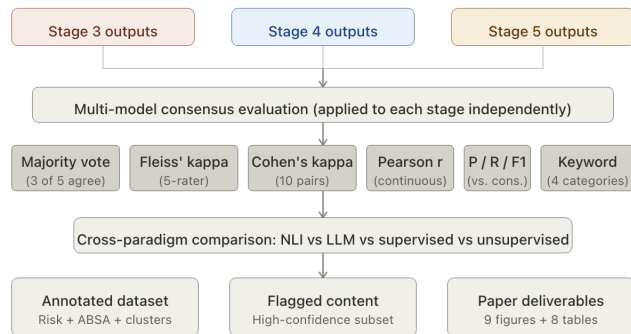


Figure 3: Six metrics (majority vote, Fleiss’ κ , Cohen’s κ , Pearson r , P/R/F1, keyword validation) enable cross-paradigm comparison. Outputs: annotated dataset, flagged subset, paper-ready figures.

ceives a structured prompt with a system message establishing the role of a misinformation detector and a user message presenting the text. We extract both the generated YES/NO response and a continuous score from the first-token logit probability $P(\text{YES})$, all in a single forward pass. GatorTron, being an encoder-only model, cannot generate text, so we implement an embedding-based approach: cosine similarity between the input text and reference sentences representing misinformation versus genuine discourse.

Due to inference time constraints, this stage processes a stratified random sample of 2,000 texts rather than the full dataset. This sample size is standard in NLP evaluation studies and provides sufficient statistical power for inter-model comparison, as seen in Figure 3.

Consensus and Agreement Metrics

A unified evaluation framework is applied to Stages 3, 4, and 5, making cross-paradigm comparison possible. The central mechanism is majority consensus: a text is flagged when three or more of five models agree. This consensus label serves as a pseudo-ground-truth for computing per-model performance metrics.

We report four agreement metrics. Fleiss’ Kappa quantifies the degree of agreement among all five models simultaneously, correcting for chance. Pairwise Cohen’s Kappa is computed for all ten model pairs, revealing which pairs converge and which diverge. Pearson correlation captures agreement in continuous score magnitude before thresholding. Per-model precision, recall, and F1 are computed against the consensus label, showing how closely each individual model aligns with the ensemble judgment.

As an external validation mechanism independent of the models themselves, we define four misinformation keyword categories (Cure Claims, Anti-Medical Rhetoric, Pseudoscience, Commercial Promotion) and measure the overlap between keyword-matched texts and model-flagged texts. If the models are capturing genuine misinformation patterns rather than spurious linguistic features, we would expect substantially higher flagging rates among keyword-matched texts than among the general population.

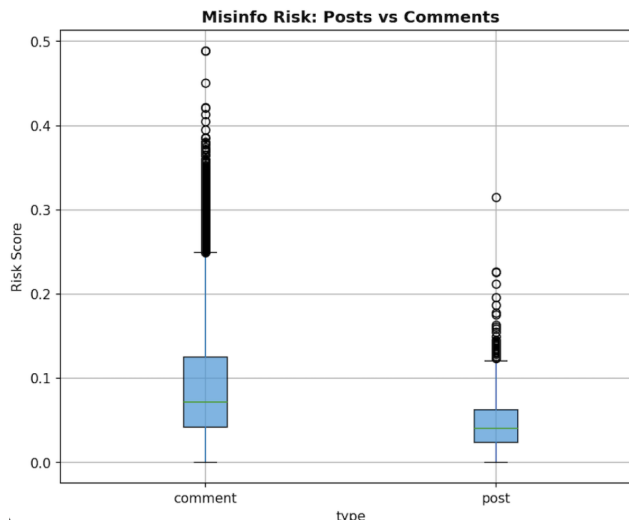


Figure 4: HSMI risk by post type. Comments show higher median risk (0.075 vs. 0.04) and more extreme outliers (max 0.49 vs. 0.32).

Results and Discussion

Dataset and Keyword Analysis

The preprocessed corpus comprises 9,324 texts (963 posts, 8,361 comments). The top unigrams are “flare” (2,455), “help” (1,814), and “year” (1,625). Bigrams reveal clinical concerns: “flare ups” (380), “hair removal” (172), “open wound” (171). Trigrams surface treatment patterns: “laser hair removal” (144), “diaper rash cream” (41), “tea tree oil” (33). TF-IDF confirms “flare” (0.057) and “help” (0.048) as the most distinctive terms. The most discussed treatments are zinc (562), surgery (555), and antibiotic (555); top symptoms are flare (2,455), wound (661), and boil (629).

Stage 1: ABSA Results

The three most prevalent aspects are “doctor and health-care experience” (2,590, 27.8%), “sharing personal experience” (2,081, 22.3%), and “pain and physical symptoms” (1,827, 19.6%), together accounting for 70% of texts. “Mental health and emotional wellbeing” appears in only 116 texts (1.2%), despite HS’s well-documented psychosocial burden.

Among emotions, “frustrated” leads affective categories (1,400, 15.0%), followed by “grateful” (834, 8.9%) and “hopeful” (604, 6.5%). Sentiment skews positive: neutral 4,447 (47.7%), positive 3,512 (37.7%), negative 1,365 (14.6%), reflecting the community’s supportive tone, as seen in Figure 5.

Stage 2: Supervised RoBERTa Classification

Seven of ten fine-tuned models produced nonzero F1. Three (DeBERTa-v3, Longformer, Clinical-Longformer) failed entirely, likely due to mismatch between their design (long documents, disentangled attention) and the short, informal Reddit texts. RoBERTa-large achieves the best F1

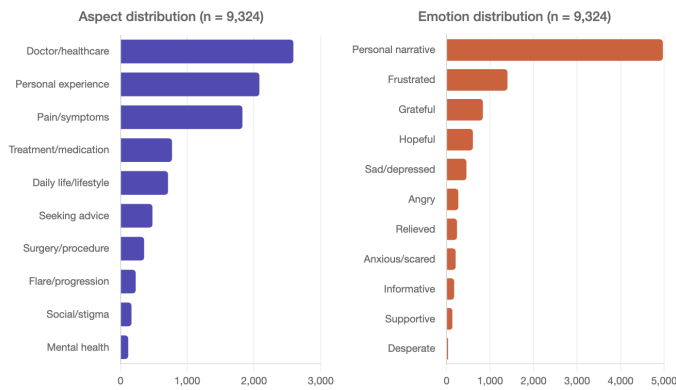


Figure 5: Aspect and emotion distributions (n=9,324). “Doctor/healthcare” dominates aspects (27.8%), while “mental health” accounts for only 1.2%. “Frustrated” is the leading affective emotion (15.0%).

of 0.7742 (AUC = 0.9405), followed by Twitter-RoBERTa (0.7083, AUC = 0.9329) and DistilRoBERTa (0.7000, AUC = 0.9295). DistilRoBERTa’s strong showing despite only 82M parameters makes it the best efficiency-performance tradeoff (2.6 min training vs. 42.9 min for RoBERTa-large).

General models outperform domain-specific across all metrics: F1 (0.570 vs. 0.411), accuracy (0.703 vs. 0.513), ROC AUC (0.743 vs. 0.555). Twitter-RoBERTa’s second-place finish suggests social media pretraining helps more than biomedical or clinical pretraining for this task.

Stage 3: Unsupervised Pattern Discovery

The four zero-shot classifiers produce right-skewed ensemble scores (median 0.12). Risk thresholds are set at $P_{75} = 0.118$ (Medium) and $P_{90} = 0.186$ (High). Model agreement: 57.7% flagged by zero models, 7.2% by three, and 0.2% by all four. Pairwise correlations range from 0.037 to 0.518.

Emotion-risk cross-tabulation reveals that disgust carries the highest HSMI score (0.10), while joy carries the lowest (0.06). UMAP + HDBSCAN identifies 2 clusters: a smaller high-risk cluster (700 texts, mean risk 0.12, dominated by anger/disgust) and a larger low-risk cluster (9,000 texts, mean risk 0.08, dominated by neutral emotion). Comments carry higher risk than posts (median 0.075 vs. 0.04), though posts express more negative sentiment (61% vs. 42%).

Stage 4: Zero-Shot NLI Detection (5 Models)

Flagging rates range from 1.1% (XLM-R-XNLI) to 17.3% (DeBERTa-Zeroshot). Majority consensus flags 182 texts (2.0%); unanimous agreement covers 27 (0.3%). Mean pairwise Cohen’s $\kappa = 0.2012$ (slight-to-fair). Pearson correlations range from 0.260 to 0.599, with BART-MNLI and RoBERTa-MNLI showing the strongest continuous-score agreement.

Per-model F1 against consensus: BART-MNLI 0.562 (P=0.40, R=0.91), RoBERTa-MNLI 0.516, XLM-R-XNLI 0.497, DeBERTa-v3-NLI 0.393, DeBERTa-Zeroshot 0.194 (high recall of 0.95 but precision of only 0.10).

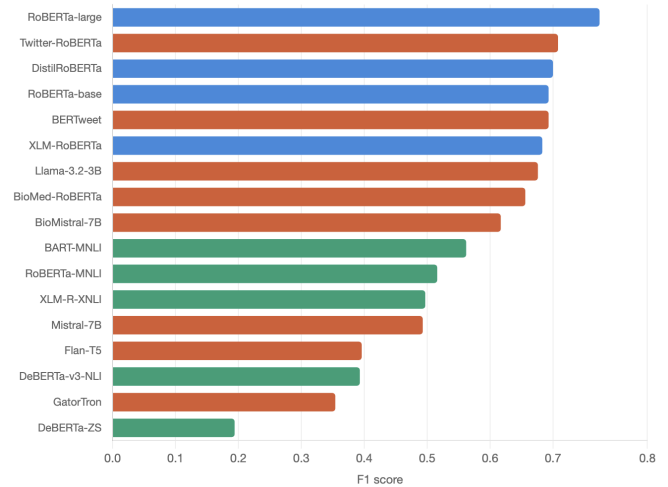


Figure 6: Cross-paradigm F1 comparison (17 models). Blue: fine-tuned RoBERTa; green: NLI classifiers; orange: LLMs. RoBERTa-large leads (0.774), Llama-3.2-3B is the top LLM (0.676), NLI models range 0.19–0.56.

Stage 5: Open-Source LLM Detection (5 Models)

On the 2,000-text sample, flagging rates diverge: Flan-T5 44.5%, GatorTron 41.6%, BioMistral 13.6%, Llama-3.2-3B 12.7%, Mistral-7B 6.2%. Flan-T5 produces near-ceiling scores ($\mu = 0.981$) with near-zero correlation to other models ($r = 0.002$ to -0.043), indicating poor calibration.

Majority consensus flags 240 texts (12.0%), six times the NLI rate. Mean Cohen’s $\kappa = 0.1551$. Llama-3.2-3B achieves the best F1 of 0.676 (balanced P=0.66, R=0.70), followed by BioMistral at 0.617. Keyword validation confirms domain relevance: Mistral shows 3.7x enrichment (21.7% keyword-matched vs. 5.9% non-keyword). GatorTron shows no enrichment (39.1% vs. 41.7%), suggesting its embedding approach does not discriminate on keyword presence. General LLMs achieve slightly higher consensus F1 (0.522 vs. 0.486) but lower flagging rates (21.2% vs. 27.6%).

Closed-Source LLM Validation

Five closed-source models classify 300 posts into 7 misinformation categories. All models identify accurate information as the majority class (46–62%). The most detected misinformation categories are trigger misinterpretation and stigma-infused narrative. Claude Sonnet 4 shows the lowest confidence (0.674 vs. 0.844–0.930 for others) but the highest sensitivity to stigma narratives (58 posts). Mean pairwise Cohen’s $\kappa = 0.62$ (moderate-to-substantial), far exceeding the open-source agreement of 0.15–0.20, as seen in Figure 8.

Cross-Paradigm Comparison

NLI classifiers are conservative (2.0% consensus rate) with higher inter-model agreement ($\kappa = 0.20$); LLMs are aggressive (12.0%) with lower agreement ($\kappa = 0.16$). Closed-source models achieve substantially higher agreement ($\kappa =$

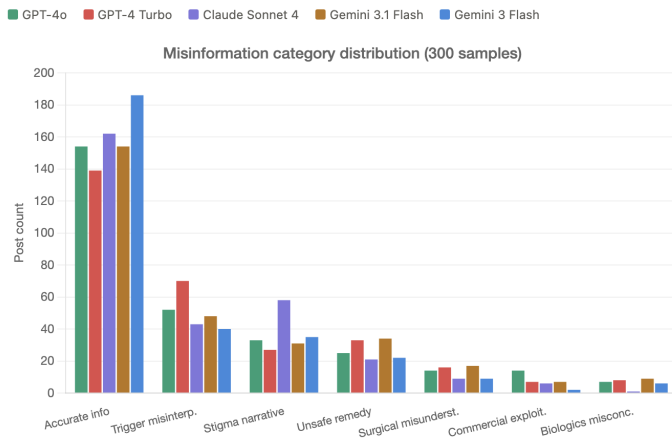


Figure 7: Misinformation category distribution across five closed-source models (300 samples).

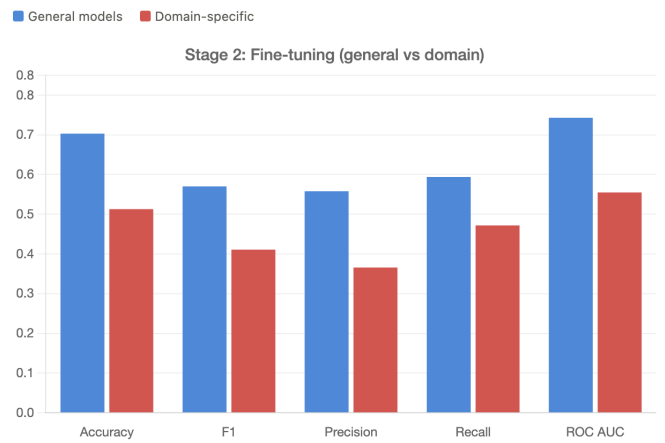


Figure 10: Stage 2 fine-tuning: general-purpose models outperform domain-specific across all metrics

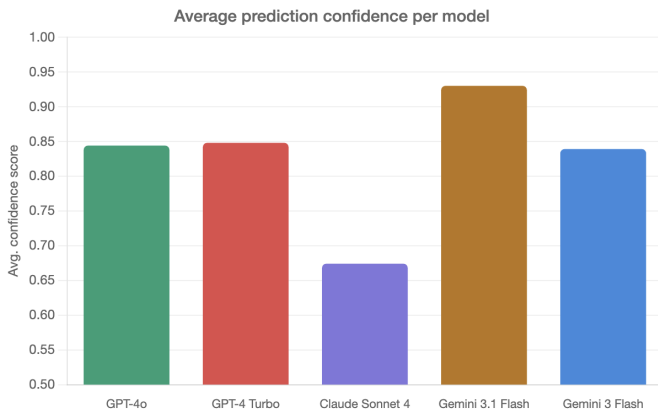


Figure 8: Average prediction confidence per closed-source model.

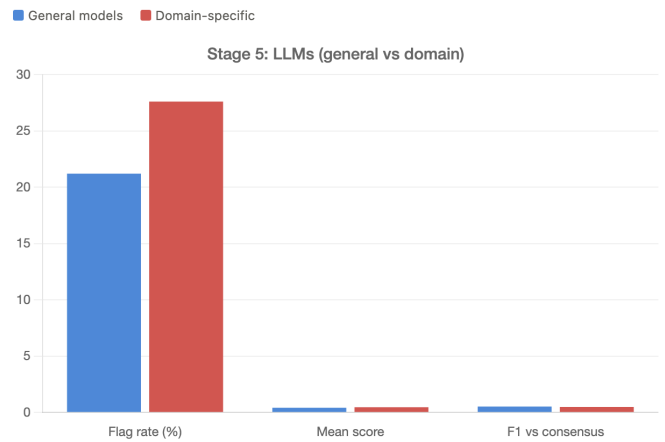


Figure 11: Stage 5 LLMs: domain-specific models flag more aggressively but achieve lower consensus F1.

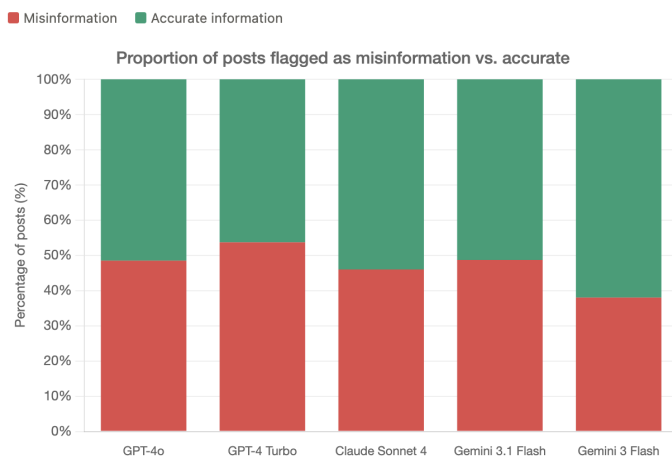


Figure 9: Proportion of posts flagged as misinformation vs. accurate

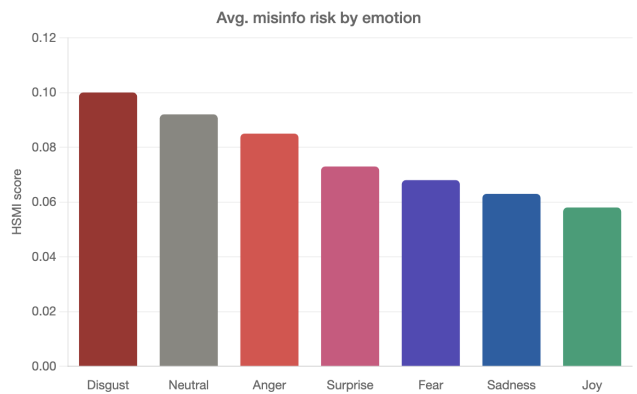


Figure 12: Corpus sentiment distribution

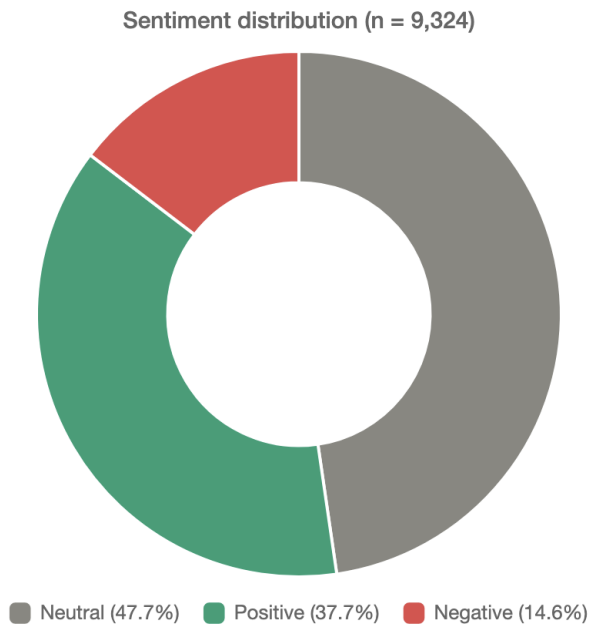


Figure 13: Average HSMI score by Ekman emotion category.

0.62) and richer 7-category classification. The best individual F1 is higher for LLMs (Llama 0.676) than NLI (BART 0.562), partly because the LLM consensus is less strict. Keyword enrichment is stronger for LLMs (up to 3.7x vs. 2x for NLI), suggesting LLMs better detect surface-level misinformation cues. Both paradigms confirm that health misinformation detection without labels remains a task where reasonable models disagree substantially, seen in Figure 6.

Overall Analysis of Results

Taken together, these results point to a few consistent patterns. Multi-model consensus, even without any human labels, is capable of surfacing misinformation signals that align with domain specific keyword heuristics and are robust across paradigms. General-purpose models outperform domain-specific ones at every stage, suggesting that the informal, conversational nature of Reddit discourse rewards broad language understanding more than narrow biomedical pretraining. The gap between open-source agreement ($\kappa = 0.15-0.20$) and closed-source agreement ($\kappa = 0.62$) indicates that model scale and instruction quality remain important factors in detection reliability. Perhaps most importantly, the emotion risk analysis reveals that misinformation in HS communities does not appear randomly it clusters around specific emotional states such as disgust and anger, and is more prevalent in comments than in original posts, pointing toward concrete moderation strategies. While no single model or paradigm solves the problem, the convergence of signals across 28 independent models provides a stronger foundation for misinformation detection than any individual classifier could offer alone, seen in Figure 10-13.

Conclusion

We present the first unsupervised framework for detecting health misinformation in Hidradenitis Suppurativa online communities, deploying 28 models across five experimental stages without requiring human annotation. Our results demonstrate that multi-model consensus serves as a viable proxy for expert labels, with general-purpose models consistently outperforming domain-specific ones and misinformation clustering around specific emotional states rather than appearing uniformly across discourse. The proposed HS Misinformation Index (HSMI) provides a continuous risk metric integrating zero-shot consensus, keyword heuristics, and sentiment signals, offering a more nuanced alternative to binary classification. While the absence of expert ground truth remains a limitation, strong alignment between model consensus and keyword validation confirms that the framework captures genuine misinformation patterns. We release the annotated dataset as the first HS misinformation benchmark and anticipate that this multi-model consensus methodology can be adapted to other underserved chronic disease communities where labeled data is scarce but detection needs are urgent.

Limitations and Future Work

This work has several limitations worth noting. The 80th-percentile threshold for pseudo-label generation assumes a fixed misinformation prevalence rather than adapting to the actual data distribution, which may inflate error rates when applied to corpora with different base rates. Exploring threshold sensitivity or calibrating against small expert-labeled subsets would better estimate true prevalence. Our misinformation taxonomy also has rough edges: categories such as commercial promotion and anti-medical rhetoric do not always correspond to factually false health claims, and anecdotal patient experiences sit in a gray area between misinformation and personal narrative, introducing label noise. A finer-grained taxonomy built with input from dermatologists would improve category precision. The scope of this study is limited to a single subreddit, and we have not tested whether these patterns hold across other platforms, languages, or chronic disease communities. All evaluation is computed against model-derived consensus rather than expert judgment, which introduces circularity. We are currently working with practicing dermatologists to annotate a representative subset with gold-standard labels, which will allow proper validation and move this work from an exploratory framework toward a reliable benchmark.

References

- Acheampong, F. A.; et al. 2020. Emotion detection in text: A review. *Information Fusion*.
- Alikhan, A.; et al. 2009. Hidradenitis suppurativa: a comprehensive review. *Journal of the American Academy of Dermatology*.
- Blei, D. M.; et al. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*.
- Bouazzi, D.; et al. 2024. Prevalence of hidradenitis suppurativa: A meta-analysis. *Dermatology*.

- Brown, T.; et al. 2020. Language models are few-shot learners. *NeurIPS*.
- Chou, W. S.; et al. 2018. Social media and health misinformation. *American Journal of Public Health*.
- Cinelli, M.; et al. 2020. The COVID-19 social media infodemic. *Scientific Reports*.
- Dalgard, F.; et al. 2015. The psychological burden of skin diseases. *Acta Dermato-Venereologica*.
- De Choudhury, M. 2014. Social media as a measurement tool for mental health. *ICWSM*.
- De Choudhury, M.; et al. 2015. Reddit and mental health discussions. *CHI*.
- Devlin, J.; et al. 2019. BERT: Pre-training of deep bidirectional transformers. In *NAACL*.
- Fox, S. 2013. The social life of health information. *Pew Research Center*.
- Frew, J. W. 2024. Unravelling the complex pathogenesis of hidradenitis suppurativa. *Journal of Investigative Dermatology*.
- Garg, A.; et al. 2017. Comorbidities of hidradenitis suppurativa. *JAMA Dermatology*.
- Guidry, J. 2017. Online health information and misinformation. *Health Communication*.
- Hay, R. J.; et al. 2014. The global burden of skin disease. *JAMA Dermatology*.
- Hutto, C. J.; and Gilbert, E. 2014. VADER: A parsimonious rule-based model for sentiment analysis. In *ICWSM*.
- Jemec, G. B. 2012. Hidradenitis suppurativa. *New England Journal of Medicine*.
- Kojima, T.; et al. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.
- Kouris, A.; et al. 2016. Quality of life and psychosocial implications in HS. *Acta Dermato-Venereologica*.
- Kouzy, R.; et al. 2020. Coronavirus goes viral: Quantifying misinformation. *Journal of Medical Internet Research*.
- Lee, J.; et al. 2020. BioBERT: A pre-trained biomedical language model. *Bioinformatics*.
- Luo, R.; et al. 2022. BioGPT: Generative pre-trained transformer for biomedical text. *Briefings in Bioinformatics*.
- Matusiak, L. 2010. Quality of life in hidradenitis suppurativa patients. *Dermatology*.
- Mohammad, S. M.; et al. 2018. SemEval-2018 Task 1: Affect in tweets. In *SemEval*.
- Moorhead, S.; et al. 2013. A new dimension of health care: systematic review of social media. *Journal of Medical Internet Research*.
- Nielsen, V. W.; et al. 2024. Genetic susceptibility to hidradenitis suppurativa and cardiometabolic disease. *JAMA Dermatology*.
- OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
- Patwa, P.; et al. 2021. Fighting an infodemic: COVID-19 fake news detection. In *EMNLP*.
- Perez-Chada, L. 2019. Social media use in psoriasis patients. *Dermatology Online Journal*.
- Ruchansky, N.; et al. 2017. CSI: A hybrid deep model for fake news detection. In *CIKM*.
- Sabat, R.; et al. 2024. Hidradenitis Suppurativa. *The Lancet*.
- Sarker, A.; et al. 2020. Self-reported COVID-19 symptoms on social media. *JMIR*.
- Saunte, D.; and Jemec, G. 2017. Hidradenitis suppurativa: advances in diagnosis and treatment. *JAMA*.
- Shen, Y.; et al. 2023. HuggingGPT: Solving AI tasks with ChatGPT and its friends. *arXiv preprint*.
- Shu, K.; et al. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explorations*.
- Singhal, K.; et al. 2023. Large language models encode clinical knowledge. *Nature*.
- Szepietowska, M.; et al. 2026. Depression and anxiety in hidradenitis suppurativa patients. *Journal of Clinical Medicine*.
- Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science*.
- Wang, W. Y. 2017. Fake news detection using deep learning. In *ACL*.
- Wang, X.; et al. 2023. Self-consistency improves chain of thought reasoning. In *ICLR*.
- Wang, Y.; et al. 2019a. Health misinformation on social media: A systematic review. *Bulletin of the WHO*.
- Wang, Y.; et al. 2019b. Social media in autoimmune diseases. *Autoimmunity Reviews*.
- Zarocostas, J. 2020. How to fight an infodemic. *The Lancet*.
- Zhou, X.; and Zafarani, R. 2020. A survey on fake news detection. *ACM Computing Surveys*.