When LLMs Disagree with Human Experts: Understanding LLM Annotation Failures in Nutrition Misinformation through Hierarchical Error Analysis using Seed Oil Narratives

Vishwaa Shah
University of North Florida
School of Computing
Jacksonville, FL, USA
N01458714@unf.edu

Indika Kahanda
University of North Florida
School of Computing
Jacksonville, FL, USA
indika.kahanda@unf.edu

Andrea Arikawa
University of North Florida
Nutrition & Dietetics
Jacksonville, FL, USA
a.arikawa@unf.edu

Abstract

Accurate linguistic annotation is crucial for creating high-quality datasets in specialized domains, yet manual labeling is often slow, expensive, and inconsistent. We present a reproducible workflow for evaluating the effectiveness of large language models (LLMs) as annotators of domain-specific health misinformation on social media. Using a data set of 169 Instagram posts on seed oils, expert nutritionists provided gold-standard labels (71% positives), which we compared against the outputs of five open-source LLMs. We introduce a hierarchical error taxonomy that categorizes LLM misclassifications according to the direction, mechanism, and contributing factors of the error, providing interpretable insights into model failures. Our analysis reveals systematic error patterns, including misinterpretation of nuanced claims and overconfidence in predictions, highlighting conditions under which LLM annotations do not align with expert judgment. Although the data set is modest in size and exhibits class imbalance, it reflects real-world distributions of nutrition-related Instagram content and motivates the need for a careful evaluation of the robustness of the LLM annotation. This study has implications for the development of frameworks for automated LLM-based annotators in the health and nutrition domains, as well as LLM developers in general.

1 Introduction

Nutrition misinformation on social media poses a growing public health challenge, shaping perceptions and behaviors around diet and health (Faruk, 2024). Platforms like Instagram, with over two billion monthly users (Dean, 2025), amplify the spread of such content due to their visual and influence-driven ecosystems. A prominent example is discourse around "seed oils," commonly consumed vegetable oils rich in unsaturated fatty acids. Although the U.S. Dietary Guidelines recommend their inclusion in a healthy diet (Dietary Guidelines for Americans, 2020), online narratives frequently portray these oils as harmful or inflammatory, linking them to chronic diseases (Petersen et al., 2024). Surveys indicate that 43% of U.S. consumers recently encountered information about seed oils primarily via social media rather than healthcare professionals or scientific sources (Balagtas and Bryant, 2025).

Given the scale and speed of misinformation, automated detection methods are increasingly important. LLMs such as GPT-4, ChatGPT, and Llama have demonstrated strong capabilities in identifying misleading health-related content (Tan et al., 2025) and can also serve as annotators, generating labels more efficiently than human experts (Tan et al., 2024; Goel et al., 2023). However, domain-specific claims, such as those regarding seed oils or n-6 PUFA intake, pose additional challenges due to complex and nuanced scientific evidence (Petersen et al., 2024).

Recent approaches improve annotation quality by combining multiple LLMs or leveraging human-in-the-loop strategies, including ensemble methods, relevancy scoring, and multi-step verification frameworks (Qiu et al., 2025; Schroeder et al., 2025). Human-LLM collaborative frameworks further enhance reliability by using model explanations to guide selective re-annotation (Wang et al., 2024; Nahum et al., 2025). In the nutrition domain, fine-tuned models like BERT and RoBERTa have successfully detected misleading content on Instagram (Lamichhane et al., 2025), highlighting the need for robust annotation pipelines that align automated detection with expert judgment (Segado-Fernández et al., 2025).

Despite these advances, systematic evaluation of LLMs as domain-specific annotators remains limited. Misperceptions about seed oils persist online, and methods for integrating LLM outputs with high-quality expert annotation are underex-

Proceedings of the 20th Linguistic Annotation Workshop (LAW XX), pages 59–74

plored. In this study, we evaluate multiple LLMs on Instagram captions about nutrition using the U.S. Dietary Guidelines as a normative reference. We compare model outputs to expert nutritionist labels and analyze annotation errors to identify common failure modes, providing actionable guidance for incorporating LLMs into annotation workflows.

While our study focuses on seed oils, it serves as a case study for evaluating LLM-based annotation workflows in a domain-specific setting. The proposed framework is designed to be adaptable to other annotation tasks, although it is empirically validated only in the nutrition misinformation domain. This framework captures linguistic and semantic nuances, claim hedging, and contextual interpretation, providing insights for designing reliable annotation processes across domains. We frame this study as an annotation error analysis, showing how evaluating human–machine disagreements can inform best practices for high-quality, domain-specific datasets. Because our operational definition of scientific consensus relies exclusively on the U.S. Dietary Guidelines Report, the resulting annotations are grounded in a U.S.-centric nutrition framework. This choice ensures consistency across expert coders but also narrows the scope of generalizability to other cultural or dietary standards.

Our contributions are threefold: (1) Present a reproducible workflow for evaluating LLMs as annotators of domain-specific health misinformation, (2) Introduce a hierarchical error taxonomy that categorizes LLM annotation failures by direction, mechanism, and contributing factors, and (3) Provide an empirical analysis of LLM annotation behavior on nutrition misinformation, identifying systematic error patterns that inform best practices for integrating LLMs into linguistic annotation pipelines.

2 Related Work

Previous work demonstrates that LLMs such as ChatGPT, GPT-4, and LLama can identify misleading health-related information with promising performance (Tan et al., 2025; Faruk, 2024; Yeung et al., 2022). These studies highlight the potential of LLMs for automated fact-checking, but their effectiveness varies across specialized topics, platforms, and cultural contexts, leaving certain domains, such as diet-specific misinformation, underexplored.

Surveys and empirical studies show that LLM-generated annotations can accelerate dataset creation, improve consistency, and reduce annotation costs (Tan et al., 2024; Goel et al., 2023; Alizadeh et al., 2025). Approaches like ensemble methods and relevancy scoring further enhance annotation reliability, mitigating heterogeneity in labeling decisions across multiple models (Qiu et al., 2025; Schroeder et al., 2025). Additionally, LLMs have been shown to detect label errors in existing datasets, improving model evaluation and downstream performance (Nahum et al., 2025).

Despite these advances, research has largely focused on generic misinformation or general annotation tasks. There remains a gap in applying LLM-based detection and annotation to domain-specific misinformation, such as dietary claims about seed oils. Our hierarchical error taxonomy extends prior work on LLM-assisted annotation by integrating direction, mechanism, and contextual contributing factors into a single framework. Unlike existing taxonomies that focus primarily on factuality or reasoning alone (Tan et al., 2024; Goel et al., 2023; Nahum et al., 2025), this structure additionally captures pragmatic and linguistic sources of misalignment such as claim hedging, evidence cues, and contextual interpretation providing a more comprehensive lens for analyzing annotation failures (Wang et al., 2024; Schroeder et al., 2025).

3 Methodology

Our methodology, shown in Figure 1, follows a four-stage pipeline. First, we collected publicly available Instagram posts discussing seed oils and preprocessed captions for textual analysis. Second, three nutrition experts independently annotated each post as *Credible* or *Misinformation* according to the 2020–2025 U.S. Dietary Guidelines, producing a gold-standard dataset. Third, five instruction-tuned LLMs were prompted using a knowledge-augmented Chain-of-Thought (CoT) approach to generate step-by-step reasoning, binary labels, and confidence scores for each post. Finally, LLM predictions were evaluated against expert labels using standard classification metrics, and misclassifications were analyzed using a hierarchical error taxonomy to identify systematic error patterns and contextual factors that affect model performance.

3.1 Data Collection

A nutrition expert identified English-language posts using Instagram's native search functionality to locate content discussing dietary fats. Only

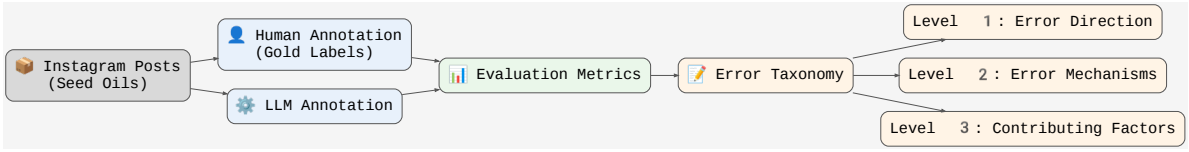

Figure 1: Full methodology pipeline from data collection to hierarchical error taxonomy analysis. The figure illustrates all major steps in our workflow, providing a visual guide to the complete processes.

posts with substantive discussion of seed oils were retained. Posts were considered relevant if captions explicitly referenced seed oils or specific seed oil products (e.g., soybean, sunflower, safflower, canola, or corn oil) in the context of nutritional or health claims. Seed oils were selected for analysis because they are frequently misrepresented on social media as harmful or inflammatory, despite being recommended sources of unsaturated fats by the U.S. Dietary Guidelines for Americans 2020–2025 (Dietary Guidelines for Americans, 2020).

To reduce personalization and geographic bias, searches were conducted using newly created, clean accounts with no prior activity, and all sessions used a VPN. This ensured retrieved posts reflected content likely visible to a new user encountering seed oil discourse. Posts were collected between May 28, 2025, and June 4, 2025, with inclusion limited to content published from January 1, 2020, through the collection period. To capture high-reach discussions, only posts from accounts with more than 5,000 followers were included.

All collected posts were publicly accessible at the time of retrieval, with no private accounts included. Usernames and identifying metadata were removed prior to analysis. Post captions and metadata were extracted using the *Apify* Instagram Post Scraper Actors[1]. The complete dataset, along with annotations and error analysis, is publicly available on Zenodo under a CC BY 4.0 License to support reproducibility and future research[2].

Our study focuses on caption text rather than image/video content to ensure privacy, feasibility, and reproducibility. Captions can be anonymized and shared with minimal ethical concerns compared to images, which may reveal identifiable individuals or sensitive health information. Instagram captions were preprocessed in Python using pandas and re library. All scripts were executed in Google Colab (Python 3.12.12) with fixed random seeds to ensure reproducibility. Preprocessing included: (a) Removing non-alphanumeric characters (excluding basic punctuation), (b) Normalizing whitespace/line breaks, and (c) Preserving emojis as separate tokens.

3.2 Human Annotation

We adopt a definition of health misinformation consistent with prior work, defined as "any health-related claim of fact that is false based on current scientific consensus" (Sylvia Chou et al., 2020). In this study, scientific consensus was operationalized according to the 2020–2025 Dietary Guidelines (Dietary Guidelines for Americans, 2020). Posts that contradicted or misrepresented these guidelines were labeled *Misinformation*; posts consistent with or not in conflict with the guidelines were labeled *Credible*.

Three expert annotators with formal training in nutrition independently evaluated each post using a detailed coding manual derived from the U.S. Dietary Guidelines for Americans 2020–2025 (see full annotation guidelines in Appendix A). Annotators were blind to each other's labels and annotated posts independently in randomized order to minimize order effects. The inter-rater reliability (IRR) among the three annotators was assessed using Fleiss' κ and the final label for each post was determined via majority vote. We treat this human consensus as the operational gold-standard for determining LLM errors. While the annotation pipeline is reproducible at the procedural level, the expert labeling step necessarily involves inter-annotator variability due to subjective judgment.

Figure 2 shows the distribution of caption lengths for misinformation and credible posts. Captions are substantial in length (overall mean: 146–166 words), requiring contextual reasoning beyond isolated claims. Misinformation posts were slightly shorter (M = 142.9, SD = 102.2) than credible posts (M = 176.9, SD = 112.9), though both categories include medium- and long-form captions exceeding

[1] https://apify.com/actors, last accessed 06/04/2025.
[2] https://doi.org/10.5281/zenodo.20141371

several hundred words.

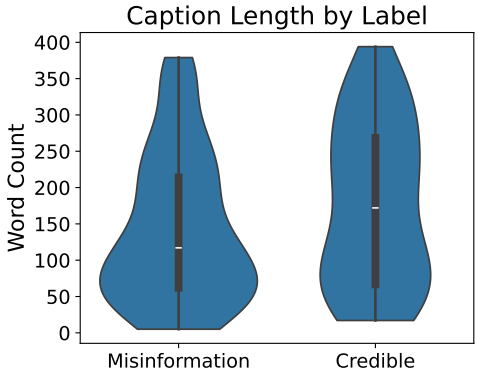

Figure 2: Caption length (word count) by expert label. Both misinformation/ credible posts contain substantial textual content, with mean lengths exceeding 140 words.

Of the 169 posts, 71% (121/169) were labeled as *Misinformation* by expert consensus, indicating class imbalance toward misinformation. All captions were manually reviewed. While some were very short or non-informative, we retained them to reflect real-world variability in Instagram captions.

3.3 LLM Annotation

3.3.1 Models

We employed five open-weight, instruction-tuned LLMs: GPT-OSS-13B, Mistral-7B-Instruct, Llama-2-7B, Zephyr-7B, and Qwen3-4B. We focus on open-weight models to ensure transparency and reproducibility of the annotation pipeline. These models span parameter scales from 4B to 13B and were selected to balance instruction-following capability and feasibility under moderate GPU constraints. Default decoding parameters were used; temperature and sampling settings were not manually overridden. Because sampling-based decoding was employed, minor stochastic variation may occur across runs.

3.3.2 Computational Environment

Preprocessing, annotation aggregation, and LLM inference were conducted in Google Colab (Linux-6.6.113+, x86_64) with approximately 15GB of GPU VRAM. Python version: 3.12.12, pandas version: 2.2.2, NumPy version: 2.0.2, re version: 2.2.1, and Ollama version: 0.17.0 were used.

LLMs were executed locally via the Ollama framework. Inference was conducted sequentially to manage memory constraints. Model outputs were parsed deterministically using regular expressions to extract structured labels and confidence scores. Inference runtime for annotating the full dataset (169 posts) varied by model size: Qwen3-4B required approximately 30 minutes; Mistral-7B, Llama-2-7B, and Zephyr-7B required 45–60 minutes; GPT-OSS-13B required 90 minutes. All models were executed sequentially.

3.3.3 Prompting Strategy

We experimented with several prompting strategies, including zero-shot, few-shot (1–5 labeled examples of credible and misinformation posts), and knowledge-augmented prompts incorporating brief excerpts from the 2020–2025 U.S. Dietary Guidelines. Across these variants, a hybrid CoT plus knowledge-augmented approach produced the most consistent and interpretable outputs, providing step-by-step reasoning that closely aligned with expert judgment.

Although we explored these alternatives during pilot development, the goal of this study was not prompt optimization but analysis of LLM annotation behavior under a fixed and interpretable configuration. All models were therefore evaluated using the same final prompting setup to ensure comparability across systems.

In our final setup, each model was conditioned as a registered dietitian and instructed to: (1) Generate step-by-step reasoning explaining its label, (2) Output a binary label (0 = credible, 1 = misinformation), and (3) Provide a confidence score, range [0.0, 1.0].

Embedding guideline excerpts in the prompt reduced normative drift and encouraged guideline-grounded reasoning. This hybrid strategy enabled transparent evaluation of LLM annotation behavior and supported systematic error analysis. The full prompt template appears in Appendix B. Outputs were parsed using Python's `re` library to extract numeric labels and confidence scores. Reasoning text was preserved for qualitative error analysis but did not influence label extraction.

3.3.4 LLM Accuracy Evaluation Protocol

Model label predictions were evaluated against expert consensus using: Accuracy, Precision, Recall, Macro-averaged F1, and Cohen's κ. Inter-LLM agreement was measured using Fleiss' κ. The LLM–human disagreement rate is computed by comparing the aggregated LLM prediction (majority vote across the five models) against the final expert consensus label for each post.

3.4 Error Taxonomy

We developed a hierarchical error taxonomy to characterize misclassifications of Instagram posts about seed oils (full taxonomy is provided in Appendix C). The taxonomy has three levels: (1) error direction, indicating whether a model over- or under-flagged misinformation; (2) error mechanism, describing broad types of model mistakes; and (3) contributing factors, capturing contextual or post-level features that may have influenced the misclassification.

The hierarchical taxonomy captures not only factual correctness but also linguistic and semantic nuances, such as claim hedging, pragmatic interpretation, and subtle contextual cues. These features reflect core challenges in linguistic annotation and provide structured insights into error behavior within the nutrition misinformation domain.

3.4.1 Level 1: Error Direction

Error direction distinguishes between **False Positives (FP):** The model labeled a credible post as misinformation, and **False Negatives (FN):** The model labeled a misinformation post as credible. This level captures tendencies toward over-flagging versus permissive behavior, which are relevant for real-world deployment.

3.4.2 Level 2: Error Mechanisms

Each misclassified post was assigned exactly one primary Level-2 error category according to a predefined coding manual (see Appendix C). Level-2 categories were defined as mutually exclusive error mechanisms. Annotators followed a hierarchical decision procedure to identify the dominant source of failure. First, errors were evaluated for Content Interpretation Errors (A), defined as cases where the model misread or failed to correctly capture the meaning or implied claims of the post. If the post was correctly interpreted but the model applied incorrect logical, causal, or evaluative reasoning, the error was coded as a Reasoning Error (B). If the reasoning process was coherent but the decision relied on incorrect, missing, or outdated domain-specific nutrition knowledge, it was coded as a Factual Knowledge Error (C). Linguistic and Stylistic Biases (D) (e.g., tone, framing, or presentation effects influencing classification) were included in the taxonomy but were not observed in the present dataset. When multiple mechanisms were present, annotators selected the dominant cause of the final misclassification based on the coding guidelines. Ambiguous cases were resolved during calibration sessions.

3.4.3 Level 3: Contributing Factors

For each error, the coders also annotated contributing factors at the post level.

Evidence Type: Indicates whether and how supporting evidence is provided in the post.

- **NONE:** No sources or supporting evidence.
- **CITED:** External sources such as articles, studies, or links are referenced.
- **DATA:** Quantitative evidence such as statistics, charts, or numerical results are presented.

Claim Strength: Captures the degree of certainty expressed in the claim.

- **ABSOLUTE:** Claims expressed with strong certainty or definitive language.
- **QUALIFIED:** Claims framed with tentative or cautious wording.
- **OPINION:** Statements reflecting personal beliefs, preferences, or subjective judgments.

Additional automated characteristics were computed to support exploratory analysis.

Textual Technicality (Flesch Reading Ease): The Flesch Reading Ease (FRE) score was calculated using the `textstat` Python package. FRE quantifies readability based on sentence length and word syllables, with higher scores indicating easier-to-read text. Posts were categorized into three levels based on the FRE score:

- **LAY** (≥ 60): Texts that are simple and easily understood by a general audience.
- **MIXED** (40–59): Moderate complexity.
- **TECHNICAL** (< 40): Texts that are dense, highly technical, or difficult to read.

Post Length: Measured in characters using Python, excluding spaces, punctuation, and special characters. Posts were categorized into:

- **SHORT** (< 100 chars):Very brief posts.
- **MEDIUM** (100–500 chars): Moderate length.
- **LONG** (> 500 chars): Extended posts.

3.4.4 LLM Errors Coding Procedure

Two coders independently annotated a 20% random subset of misclassified posts. Inter-rater reliability was computed using Cohen's κ. Coders then

met to resolve disagreements and clarify category definitions, with an expert guiding calibration to ensure consistent interpretation. The remaining errors were annotated by a single trained coder using the agreed-upon taxonomy, revisiting ambiguous cases as needed. The full coding manual is provided in Appendix C.

3.4.5 Error Multi-Factor Analysis

We first used Level 1 to identify all misclassified predictions. Level 2 then characterized the type of error for each model (content interpretation, reasoning, factual knowledge), while Level 3 examined post-level features: claim strength, evidence type, textual style, and caption length for each misclassification. Level-3 characteristics were counted once per misclassified prediction. Combining Level-2 error mechanisms with Level-3 features allowed us to explore how content and surface level factors relate to model specific error tendencies, including FP and FN. Confidence-weighted analyses were conducted to examine how LLM self-reported model confidence correlates with error direction.

4 Results

4.1 Human Annotation Reliability

Three nutrition-trained annotators independently labeled 169 Instagram posts. Pairwise Cohen's κ ranged from 0.789 to 0.828, with Fleiss' $\kappa = 0.813$, indicating substantial agreement. The human disagreement rate (11.8%) is computed as the proportion of instances on which at least one of the three expert annotators disagreed prior to majority voting. This reflects a set of borderline cases with inherent annotation ambiguity. We include a representative case study of annotator disagreement in Appendix D, illustrating how such borderline claims map onto our hierarchical error taxonomy.

4.2 LLM Annotation Performance

4.2.1 Inter-Model Agreement

Inter-model agreement among the five LLMs was lower than human agreement, with pairwise Cohen's κ spanning 0.237 to 0.813 and Fleiss' $\kappa = 0.417$, suggesting moderate consistency. While over half of all LLM predictions (56.8%) disagreed with at least one other LLM, only 11.8% of predictions differed from the human consensus labels. As a descriptive observation, the identical magnitude of these two rates (11.8%) is notable in this dataset, although it arises from different underlying processes. This LLM–human alignment rate is comparable in magnitude to the human inter-annotator disagreement rate reported in Section 4.1, although the two measures reflect different evaluation procedures and are not directly equivalent.

This pattern indicates that LLMs generally align with human judgment on clear-cut cases, with most disagreements arising from inherently ambiguous posts. As Figure 3 shows, full agreement across all five LLMs occurred for 42% of posts, with multi-model disagreement being rare.

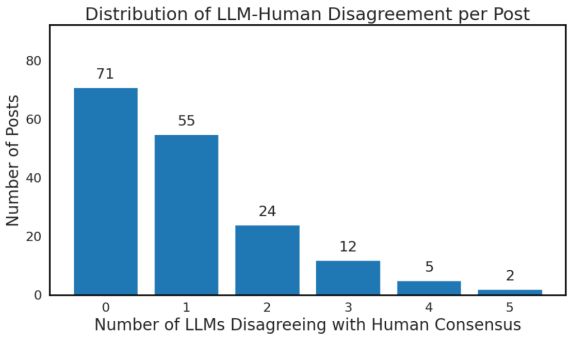

Figure 3: Distribution of the number of LLMs disagreeing with human consensus.

4.2.2 Individual LLM Performance

Table 1 reports individual model metrics against human consensus. GPT-OSS and Qwen3 achieved the highest F1 and Cohen's κ, whereas Llama-2 performed substantially worse than the other models, with notably lower F1 score and Cohen's κ, indicating weaker agreement with human annotations. Zephyr and Mistral showed intermediate performance, with Mistral achieving relatively balanced metrics across measures.

Model	Acc.	Prec.	Rec.	F1	κ
GPT-OSS	0.905	0.973	0.893	0.931	0.781
Mistral	0.852	0.864	0.942	0.901	0.609
Llama-2	0.615	0.811	0.603	0.692	0.209
Zephyr	0.746	0.933	0.694	0.796	0.476
Qwen3	0.882	0.939	0.893	0.915	0.720

Table 1: Individual LLM performance against human consensus. κ is Cohen's κ.

4.3 LLM Error Analysis

Of the 169 Instagram posts, 98 contained at least one LLM prediction that disagreed with the expert consensus. These 98 posts account for 169 of the 845 total predictions (169 posts $\times$ 5 models, 20%). For Level-2 analysis, we examine each

misclassified prediction using our hierarchical error taxonomy to systematically characterize model failure modes. Level-3 analysis considers post-level features across these 98 posts to identify content characteristics associated with LLM errors.

4.3.1 Level 1 Error Direction

Across all models, we observed 169 misclassifications, with substantially more FN (n = 118) than FP (n = 51), indicating a general tendency to under-flag misinformation. Model-level comparison shows distinct behavioral tendencies: models such as Llama-2 and Zephyr produced a higher proportion of FNs, whereas GPT-OSS and Mistral exhibited a more balanced FP/FN distribution. Figure 4 illustrates the distribution of FN and FP across models.

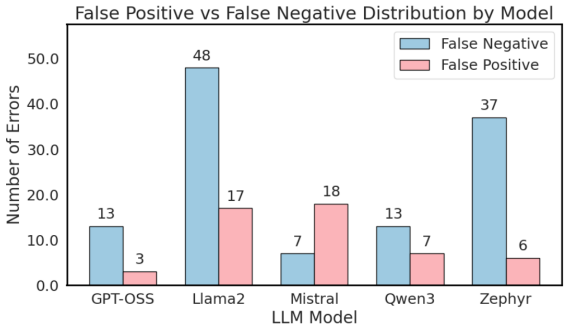

Figure 4: False positive vs. false negative distribution.

4.3.2 Inter-Coder Reliability for Errors

For a subset of misclassified posts (20%) independently annotated by two coders, inter-rater agreement was substantial, with Cohen's $\kappa = 0.774$ for primary error mechanism (Level 2) codes. For post-level contributing factors, agreement was $\kappa = 0.726$ for Evidence Type and $\kappa = 0.758$ for Claim Strength. Disagreements were resolved through discussion with an expert coder. Following this calibration, the second coder completed annotation of the remaining posts using the clarified taxonomy, ensuring consistent coding across the full dataset.

4.3.3 Level 2 Error Mechanisms

Content interpretation errors dominated LLM annotation failures, followed by factual knowledge and reasoning errors (see Figure 5). No errors were attributed to linguistic or stylistic bias in this dataset; however, our hierarchical taxonomy is designed to capture such issues in corpora where style influences annotation, making it broadly applicable.

As Figure 5 shows, most failures arise from challenges in understanding the meaning, nuance, and implied claims in social media posts, highlighting areas where LLMs require additional grounding or contextual reasoning to align with expert judgment.

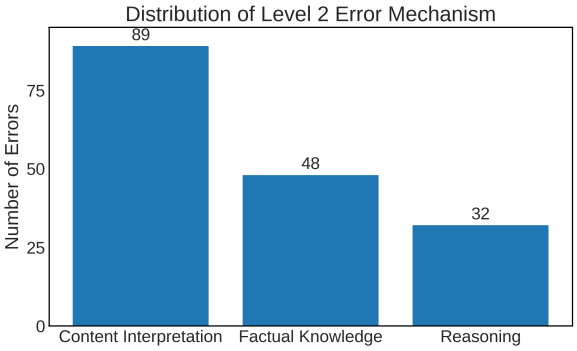

Figure 5: Distribution of Level 2 error mechanisms across all misclassified LLM predictions.

4.3.4 Level 3 Post Characteristics

Analysis of post-level features (Figure 6) shows that most posts contained QUALIFIED claims, with fewer ABSOLUTE statements or OPINION-based content. Evidence patterns show that the majority of posts contained NONE evidence, while only a small fraction included DATA or CITED sources. Textual style was primarily LAY or MIXED, with fewer TECHNICAL posts. Most captions were relatively LONG. This overview highlights the typical contexts in which LLM annotation was applied, providing insight into the types of content that pose challenges for automated labeling. All counts reported here are based on the 98 posts that contained at least one LLM prediction that disagreed with the expert consensus.

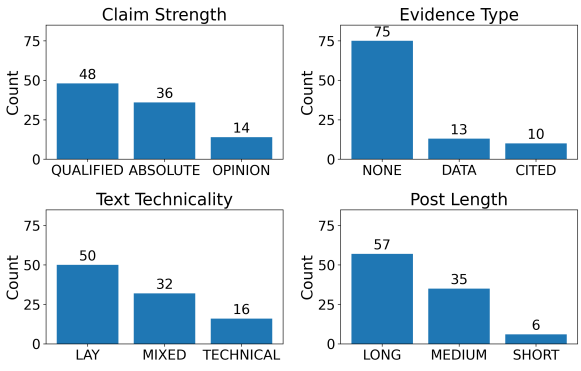

Figure 6: Distribution of Level 3 post characteristics across the dataset, including claim strength, evidence type, textual technicality, and caption length.

4.4 Multi-Factor Error Analysis

Across models, the largest number of misclassifications occurred for Llama-2 (n = 65), followed by Zephyr (n = 43) and Mistral (n = 25), while Qwen3 (n = 20) and GPT-OSS (n = 16) produced fewer errors overall. Figure 7 illustrates the breakdown of misclassifications by Level-2 error type for each model.

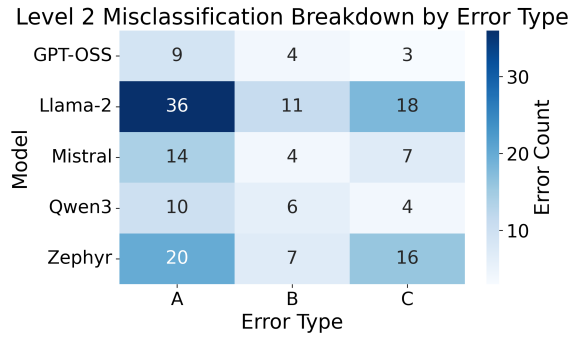

Figure 7: Breakdown of misclassifications by Level 2 error type for each model. Content interpretation errors (A), reasoning errors (B), and factual knowledge errors (C) are shown.

Analysis of Level 3 post characteristics reveals systematic associations with specific Level 2 errors. Level 3 characteristics are counted once per misclassified prediction; thus, posts with multiple misclassifications contribute multiple observations. This results in 169 misclassifications derived from 98 posts (see Appendix E for detailed heatmaps).

Content Interpretation Errors (A) were most common in posts containing ambiguous or nuanced claims. Error rates increased when claims used tentative language (QUALIFIED) or lacked supporting evidence (NONE), suggesting that models struggle with hedged or context-dependent statements. Posts written in LAY language and those with LONG captions were also more frequently misinterpreted, indicating that discourse complexity may contribute to annotation errors.

Reasoning Errors (B) occurred primarily in posts relying on anecdotal or loosely causal reasoning. These errors were especially common in posts containing QUALIFIED or OPINION-based claims, suggesting difficulty evaluating tentative or context-dependent arguments. Factual Knowledge Errors (C) were most often associated with posts requiring domain-specific nutrition knowledge. Posts lacking supporting evidence (NONE) accounted for most factual knowledge failures, whereas posts containing CITED or DATA sources were more reliably annotated. This pattern suggests that explicit evidentiary cues help anchor model judgments and reduce knowledge-related annotation errors.

Across error types, post-level characteristics consistently influenced misclassification patterns. QUALIFIED claims produced the largest number of errors overall, while ABSOLUTE claims more frequently resulted in FN errors. The absence of supporting evidence exacerbated all error types. LONG captions and LAY style language were also associated with higher error rates, likely reflecting increased discourse complexity. These results indicate that LLM annotation errors are shaped by interactions between linguistic framing, evidential context, and domain knowledge requirements.

4.5 Confidence Calibration

LLM self-reported confidence varied across models. Calibration curves in Figure 8 illustrate that GPT-OSS and Zephyr are relatively well-aligned with empirical correctness, whereas Llama-2 and Mistral exhibit over- or under-confidence. Notably, of 169 total misclassified outputs, 149 were reported with confidence ≥ 0.8, indicating systematic overconfidence.

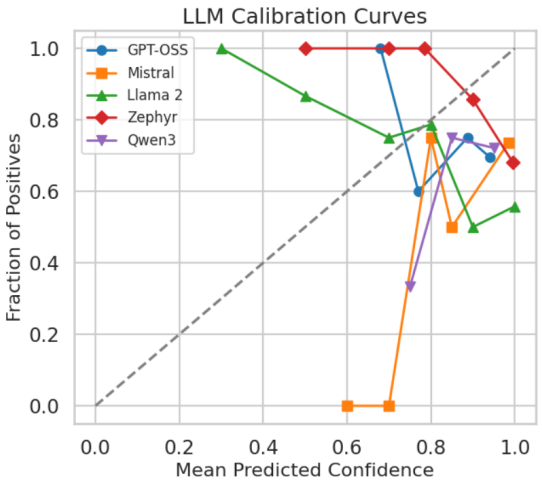

Figure 8: Calibration curves comparing predicted LLM confidence vs. empirical correctness.

5 Conclusion

We present a reproducible, expert-anchored workflow for annotating health misinformation on social media using LLMs, complemented by a hierarchical error taxonomy that categorizes annotation failures into interpretable mechanisms. Our results show that LLM errors are systematic rather than

random, arising primarily from miscalibration of claim strength, insufficient grounding in evidence, and challenges in interpreting hedged or absolute language. Persistent overconfidence in incorrect predictions indicates that failures are driven by pragmatic interpretation and epistemic misalignment, rather than surface-level textual complexity. The taxonomy captures linguistic and semantic nuances, such as claim hedging and subtle contextual cues, highlighting challenges that extend beyond nutrition misinformation and informing general best practices in linguistic annotation.

By linking error mechanisms to post characteristics and model behavior, this workflow provides guidance for when LLMs are reliable and for best practices in LLM-based annotation, including knowledge-augmented prompting and step-by-step reasoning aligned with expert judgment. It also supports transparent analysis of human–machine disagreement, enabling systematic evaluation of annotation quality and illuminating recurring failure modes. Overall, this study provides a reproducible and linguistically grounded insight for evaluating LLM annotations, offering insight for improving annotation reliability in domain-specific NLP tasks.

Future work should scale the workflow to larger and more diverse corpora, explore multimodal reasoning, and validate taxonomy coding with multiple annotators to strengthen reliability and generalizability. Future work may also incorporate active-learning loops in which LLM explanations trigger selective re-annotation, and evaluate ensemble disagreement as an uncertainty signal. These directions would deepen understanding of both annotation reliability and model-driven quality control strategies in domain-specific contexts.

Limitations

This study has several limitations. The dataset is modest in size (169 posts), which may limit the generalizability of the quantitative findings. The study focuses on expert-driven annotation and qualitative error analysis within a specific domain of nutrition misinformation rather than large-scale model evaluation. Additionally, the dataset is topic-specific ("seed oils"), English-only, and grounded in U.S. dietary guidelines, which further limits generalizability across domains, cultures, and platforms. Posts were restricted to accounts with more than 5,000 followers to prioritize high-reach content that is more likely to influence public discourse. However, this restriction may introduce sampling bias and may not fully capture misinformation circulating in smaller or private communities.

The annotation protocol relies on three expert annotators and majority voting to determine gold labels. Although inter-rater reliability was substantial (Fleiss' $\kappa = 0.813$), the disagreement rate indicates that some posts are inherently ambiguous, and the gold labels should be interpreted as consensus judgments rather than absolute truth.

Although binary labeling aligns with prior misinformation detection work and enables consistent evaluation, health-related claims often exist on a spectrum rather than as strictly true or false categories. Posts that combine accurate information with misleading interpretations may be especially difficult to annotate. Future work could explore ordinal or multi-label annotation schemes to better capture such nuance and annotator uncertainty.

The study evaluates a limited set of five open-weight language models under a single prompting configuration. Model outputs may vary with different prompts, decoding parameters, or repeated runs due to stochasticity, and the analysis does not include larger proprietary or multimodal models, which may behave differently. The error taxonomy involves interpretive coding, and although inter-coder reliability was measured and calibration performed, some subjectivity is inherent in categorizing errors.

Future work may extend this evaluation to larger proprietary or multimodal models to examine whether similar annotation error patterns emerge across different model architectures. Finally, the analysis focuses exclusively on caption text, whereas Instagram is inherently multimodal; visual content may influence how claims are interpreted, so these findings pertain specifically to textual linguistic annotation.

Ethical Considerations

This study uses only publicly accessible Instagram posts and complies with platform policies and institutional research standards. No private accounts were accessed, and all identifying information was removed prior to analysis. Any released dataset will contain anonymized caption text only.

We acknowledge that misinformation detection systems may produce false positives or false negatives with uneven social consequences. The models evaluated here are intended for research purposes

and should not be deployed for fully automated moderation without human oversight. We discourage applications involving surveillance, censorship, or suppression of legitimate expression.

Our goal is to advance transparency and reproducibility in health misinformation annotation research while prioritizing privacy, fairness, and responsible use.

Acknowledgments

We thank Charlotte Martin, Asal Abbaszadeh, Richard Loftis, and Alan Flanagan for their valuable contributions to extracting and annotating the original posts. The authors also gratefully acknowledge the School of Computing, College of Computing, College of Computing, Engineering and Construction, and Graduate School at University of North Florida for their support and funding, which made this research possible.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2025. Open-source llms for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1):17.

Joseph Balagtas and Elijah Bryant. 2025. Consumer food insights. Technical report, Center for Food Demand Analysis and Sustainability, Purdue University.

Brian Dean. 2025. Instagram demographic statistics: How many people use instagram in 2024? Accessed: 2025-04-25.

Dietary Guidelines for Americans. 2020. Dietary guidelines for americans, 2020-2025 and online materials. https://www.dietaryguidelines.gov/resources/2020-2025-dietary-guidelines-online-materials.

Tanjim Bin Faruk. 2024. Evaluating the performance of large language models in scientific claim detection and classification. https://arxiv.org/abs/2412.16486.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, and 1 others. 2023. Llms accelerate annotation for medical information extraction. In *machine learning for health (ML4H)*, pages 82–100. PMLR.

Prajwol Lamichhane, Indika Kahanda, Andrea Arikawa, Charlotte Martin, Maribel Garcia, Camila Figueiredo, and Haivan Benjamin. 2025. Exploring the feasibility of identifying nutrition misinformation on social media. In *Proceedings of the ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, pages 319–323.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. 2025. Are llms better than reported? detecting label errors and mitigating their effect on model performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26770–26797.

Kristina S Petersen, Kevin C Maki, Philip C Calder, Martha A Belury, Mark Messina, Carol F Kirkpatrick, and William S Harris. 2024. Perspective on the health effects of unsaturated fatty acids and commonly consumed plant oils high in unsaturated fat. *British Journal of Nutrition*, pages 1–12.

Jiaxing Qiu, Dongliang Guo, Natalie Papini, Noelle Peace, Hannah F Fitterman-Harris, Cheri A Levinson, Tom Hartvigsen, and Teague R Henry. 2025. Labeling free-text data using language model ensembles. *arXiv preprint arXiv:2501.08413*.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? investigating llm-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25771–25795.

Sergio Segado-Fernández, Beatriz Jiménez-Gómez, Pedro Jesús Jiménez-Hidalg, María del Carmen Lozano-Estevan, and Iván Herrera-Peco. 2025. Disinformation about diet and nutrition on social networks: a review of the literature. *Nutricion hospitalaria*, 42(2).

Wen-Ying Sylvia Chou, Anna Gaysynsky, and Joseph N Cappella. 2020. Where we go from here: health misinformation on social media.

Dongmei Tan, Yi Huang, Ming Liu, Ziyu Li, Xiaoqian Wu, and Cheng Huang. 2025. Identification of online health information using large pretrained language models: Mixed methods study. *Journal of Medical Internet Research*, 27:e70733.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, pages 1–21.

Andy Wai Kan Yeung, Anela Tosevska, Elisabeth Klager, Fabian Eibensteiner, Christos Tsagkaris, Emil D Parvanov, Faisal A Nawaz, Sabine Völkl-Kernstock, Eva Schaden, Maria Kletecka-Pulker, and 1 others. 2022. Medical and health-related misinformation on social media: bibliometric study of the scientific literature. *Journal of Medical Internet Research*, 24(1):e28152.

Appendix

A Annotation Guidelines

A.1 Purpose of These Guidelines

These guidelines ensure that all human coders classify social media nutrition posts consistently using:

- A two-level accuracy scale (0 or 1)
- Rules aligned with the *Dietary Guidelines for Americans, 2020–2025 (DGA)* and the *2020 DGAC Scientific Report*

Coders should reference only the content presented in the post and apply the rules exactly as described.

A.2 What Coders Evaluate

For each post, coders assign:

1. Accuracy classification (1 or 0)
2. An optional brief justification (2–3 sentences)

Coders do **not** evaluate:

- Creator's biography
- Hashtags
- Engagement metrics
- Intent or motivation
- User comments
- Information not shown in the post itself

A.3 Evidence Standard

Coders must evaluate claims solely against:

- *Dietary Guidelines for Americans 2020–2025*
- *DGAC Scientific Report (2020)*

Coders should **not** use:

- Personal knowledge
- Other nutrition guidelines
- Mechanistic physiology not discussed in the DGA/DGAC
- External sources, papers, or reputable organizations
- Internet searches

A.4 The Two-Level Accuracy Classification

1 - Accurate or Mostly Accurate A post is **Accurate** when:

- All nutrition claims fully align with DGA/DGAC
- Any minor inaccuracies do not change the overall message

Examples of Accurate Content

- Saying omega-6 fats are not inflammatory
- Correcting misinformation about seed oils

A post is **Mostly Accurate** when:

- It is evidence-aligned overall
- It contains some non-trivial inaccuracies, but the main message remains correct

Typical patterns

- Slight exaggeration without reversing the evidence

0 - Mostly Inaccurate or Inaccurate A post is **Mostly Inaccurate** when:

- It contains some correct information
- Misleading or incorrect claims dominate

Typical patterns

- Mixing evidence-based advice with unsupported mechanistic claims
- Overstating inflammation, hormone effects, oxidation, or fear-based messages about oils
- Encouraging elimination of seed oils while acknowledging they contain essential fats

A post is **Inaccurate** when:

- The core message contradicts DGA/DGAC
- Accurate statements, if present, are trivial

Examples

- "Butter is heart healthy; seed oils cause disease."
- "Seed oils are toxic poisons that cause inflammation."

A.5 Reasoning Requirements (2–3 Sentences)

Coders may provide a short justification for doubtful annotations. The justification must:

- Evaluate the overall message rather than line-by-line details
- Reference only areas covered by the DGA/DGAC
- Not mention excluded content (e.g., "the DGA doesn't cover . . . ")
- Avoid SEO phrases or quoting the creator directly
- Clearly explain why the classification was chosen

A.6 Special Situations

Mixed-accuracy posts Code based on the **overall message**, not on isolated statements.

Mechanistic claims (inflammation, oxidation, hormones) If the mechanism is not discussed in the DGA:

- Ignore the mechanism itself
- Evaluate whether the dietary conclusion contradicts DGA guidance

Reactive content If a post stitches another creator's video:

- Ignore the introductory clip
- Code only the creator's commentary

B Full Prompt Template

Prompt Template Used for All Models

> You are a registered dietitian and an expert in evaluating health-related claims. Can you identify whether the following Instagram post contains misinformation that contradicts the Dietary Guidelines for Americans 2020–2025?
>
> **Key points from the guidelines:**
>
> - Oils provide essential fatty acids and are part of a healthy diet.
> - Tropical oils like coconut and palm oils are high in saturated fat.
> - Replace high saturated fat foods with vegetable oils.
>
> Post: {post}
>
> Provide step-by-step reasoning.
>
> End with:
>
> Label: 0 if credible, 1 if misinformation
> Confidence: 0.0 to 1.0

C Error Taxonomy Coding Manual

C.1 Instructions for Coders

This manual guides the classification of errors made by LLMs when annotating Instagram posts about seed oils as *misinformation* or *credible*.

Before coding:

- Familiarize yourself with all error categories and examples.
- Each error should receive **one primary error type**.
- Up to **two secondary contributing categories** may be assigned when appropriate.
- When uncertainty arises, discuss the case with the other coder.

C.2 Coding Process

1. Read the Instagram post carefully.
2. Review the **ground truth label** determined through human consensus.
3. Note the **LLM's predicted label**.
4. Identify why the LLM made the incorrect classification.
5. Assign the appropriate error category and any contributing factors.

C.3 Level 1: Error Direction

Errors are first coded by direction:

- **FP (False Positive)**: The LLM labeled the post as misinformation, but the post is credible.
- **FN (False Negative)**: The LLM labeled the post as credible, but the post contains misinformation.

C.4 Level 2: Error Mechanism Categories

A. Content Interpretation Errors

These errors occur when the LLM misinterprets the meaning or intent of the content.

Nuance Blindness: Failure to detect linguistic cues such as sarcasm, irony, or hedging that alter the meaning of the post.

- "Yeah sure, seed oils are definitely going to kill us all "
- "Love how everyone became a biochemist overnight to explain seed oils"

Context Insensitivity: Failure to incorporate contextual signals that influence interpretation.

- Ignoring hashtags such as #sarcasm or #justkidding.
- Treating posts from expert accounts the same as unverified sources.

Claim Granularity Errors: Failure to distinguish between neutral factual statements, nuanced claims, and extreme assertions.

- "Seed oils contain omega-6 fatty acids" labeled as misinformation.
- "I avoid seed oils" treated as a factual health claim rather than a personal preference.

Source Credibility Misjudgment: Incorrect evaluation of cited sources or reliance on presentation style rather than evidence quality.

- Credible health organization quoted but ignored by the model.
- Misinformation written in academic style treated as credible.

B. Reasoning Errors

These errors arise from flawed reasoning processes when evaluating claims.

Over-Generalization: Applying overly broad patterns instead of evaluating the specific claim.

- Automatically labeling any criticism of seed oils as misinformation.
- Pattern matching: "seed oils" + "inflammatory" leading to automatic misclassification.

Under-Generalization: Accepting problematic claims as credible because they are framed cautiously.

- "In my experience seed oils caused my health problems" treated as credible evidence.
- "I'm just asking questions about seed oils" interpreted as neutral inquiry.

Logical Fallacy Blindness: Failure to detect flawed reasoning such as anecdotal evidence or correlation–causation confusion.

- "I stopped eating seed oils and my acne disappeared, therefore seed oils cause acne."
- Anecdotal claims treated as scientific evidence.

C. Factual Knowledge Errors

These occur when the LLM lacks correct domain knowledge or invents information.

Outdated Knowledge: Reliance on obsolete information or outdated scientific consensus.

- Referencing outdated dietary guidance rather than more recent recommendations.

Factual Hallucination: Confidently asserting incorrect facts or nonexistent evidence.

- Claiming a specific study proves seed oils are toxic when no such study exists.

Domain Knowledge Gaps: Insufficient understanding of nutrition or biochemical concepts required to evaluate claims.

- Confusion between omega-3 and omega-6 fatty acids.
- Misinterpreting oxidative stress mechanisms.

D. Linguistic and Stylistic Biases

These errors occur when stylistic features influence classification rather than factual accuracy.

Formality Bias: Judgments influenced by writing style instead of claim validity.

- Academic-sounding misinformation labeled as credible.
- Accurate information written casually labeled as misinformation.

Confidence Confusion: Confusing the author's confidence level with factual correctness.

- Highly confident misinformation treated as credible.
- Cautious scientific language interpreted as uncertainty or misinformation.

Emotional Language Sensitivity: Overreacting to emotional tone rather than evaluating factual content.

- Fearful language triggering a misinformation label even when the claim is accurate.

C.5 Contributing Factors

For each error, coders may optionally record contextual factors that help explain the misclassification.

- **Evidence Provided**
  - **NONE**: No sources or supporting evidence are provided.
  - **CITED**: External sources (e.g., articles, studies, organizations) are referenced.
  - **DATA**: Quantitative evidence such as statistics, charts, or graphs is presented.
- **Claim Strength**
  - **ABSOLUTE**: Strong certainty (e.g., "always", "never", "proves").
  - **QUALIFIED**: Tentative or cautious wording (e.g., "may", "might", "suggests").
  - **OPINION**: Personal beliefs or preferences (e.g., "I think", "in my experience").

C.6 Inter-Rater Reliability Protocol

To ensure consistent application of the hierarchical error taxonomy, the following procedure should be followed:

1. Two coders independently annotate a randomly selected **20% subset** of the dataset using the hierarchical error taxonomy described above.
2. After both coders complete their annotations, inter-rater reliability should be calculated using **Cohen's** κ.
3. The coders then meet to review cases of disagreement and discuss the reasoning behind their coding decisions.
4. During this discussion, an expert coder clarifies category definitions and resolves ambiguities in the taxonomy to ensure that both coders interpret the categories consistently.
5. After this calibration discussion, the remaining errors in the dataset are annotated by the trained coder using the agreed-upon interpretations of the taxonomy.

C.7 Illustrative Examples

Example 1

Post: "So apparently seed oils are the reason I'm tired all the time? Sure Jan"

Ground Truth: Credible

LLM Label: Misinformation

Error Direction: FP

Error Mechanism: A. Content Interpretation Errors

Explanation: LLM missed sarcasm ("Sure Jan" meme reference)

Contributing Factors:
- Length: SHORT
- Technicality: LAY
- Evidence: NONE
- Claim Strength: ABSOLUTE

Example 2

Post: "I stopped using seed oils 6 months ago and my inflammation completely disappeared. Do your own research!"

Ground Truth: Misinformation

LLM Label: Credible

Error Direction: FN

Error Mechanism: B. Reasoning Errors

Explanation: LLM failed to identify correlation/causation fallacy and accepted anecdotal evidence as proof

Contributing Factors:
- Length: MEDIUM
- Technicality: MIXED
- Evidence: ANECDOTAL
- Claim Strength: ABSOLUTE

D Human Disagreement Case Study

To illustrate sources of annotation ambiguity and how they relate to our hierarchical error taxonomy, we present a representative case from the subset of posts where expert annotators initially disagreed prior to adjudication.

D.1 Post Example

> "Plot twist not all oils are created equal! Say goodbye to those heavily processed seed oils that are lurking in your pantry. Your body (and your tastebuds) will thank you for choosing whole food alternatives like olive oil coconut oil and avocado oil instead instead. Who else is joining the real food revolution? RealFood HealthyChoices NutritionTips SeedOils HealthySwaps WellnessJourney"

D.2 Human Annotation Disagreement

This post was independently annotated by three nutrition experts. One annotator labeled the post as **Credible**, interpreting the content as a general dietary suggestion promoting whole food alternatives without making an explicit false factual claim. In contrast, two annotators labeled the post as **Misinformation**, focusing on the implied negative framing of seed oils as "heavily processed" and the absence of scientific evidence supporting the suggested health inference.

Following majority voting, the final gold label for this instance is **Misinformation**.

D.3 LLM Prediction Error

Among the evaluated models, only Llama-2 incorrectly classified this post as **Credible**, constituting a False Negative (FN) with respect to the human consensus label.

D.4 Taxonomy-Based Analysis

Using our hierarchical error taxonomy, this case is categorized as follows:

Level 1 (Error Direction): False Negative (FN)

Level 2 (Error Mechanism): Content Interpretation Error

The misclassification arises from divergent interpretation of the post's implied claims. While the

model appears to focus on the absence of explicit factual assertions, human annotators considered the implicit causal framing (i.e., seed oils as harmful) as sufficient for a misinformation label.

Level 3 (Contributing Factors):

- **Claim Strength:** QUALIFIED / IMPLIED ABSOLUTE
- **Evidence Type:** NONE
- **Textual Technicality:** LAY
- **Post Length:** Short

D.5 Discussion

This example highlights a common source of disagreement in nutrition-related misinformation annotation: posts that combine factual product substitutions with implicit health claims. Although no explicit scientific claim is made, the framing of seed oils as "heavily processed" and the endorsement of alternative oils introduces an evaluative inference that leads to divergence in human interpretation. This case illustrates how Content Interpretation Errors in LLMs often stem from failures to capture pragmatic implications rather than explicit factual inaccuracies.

E Level 2 and Level 3 Error Heatmaps

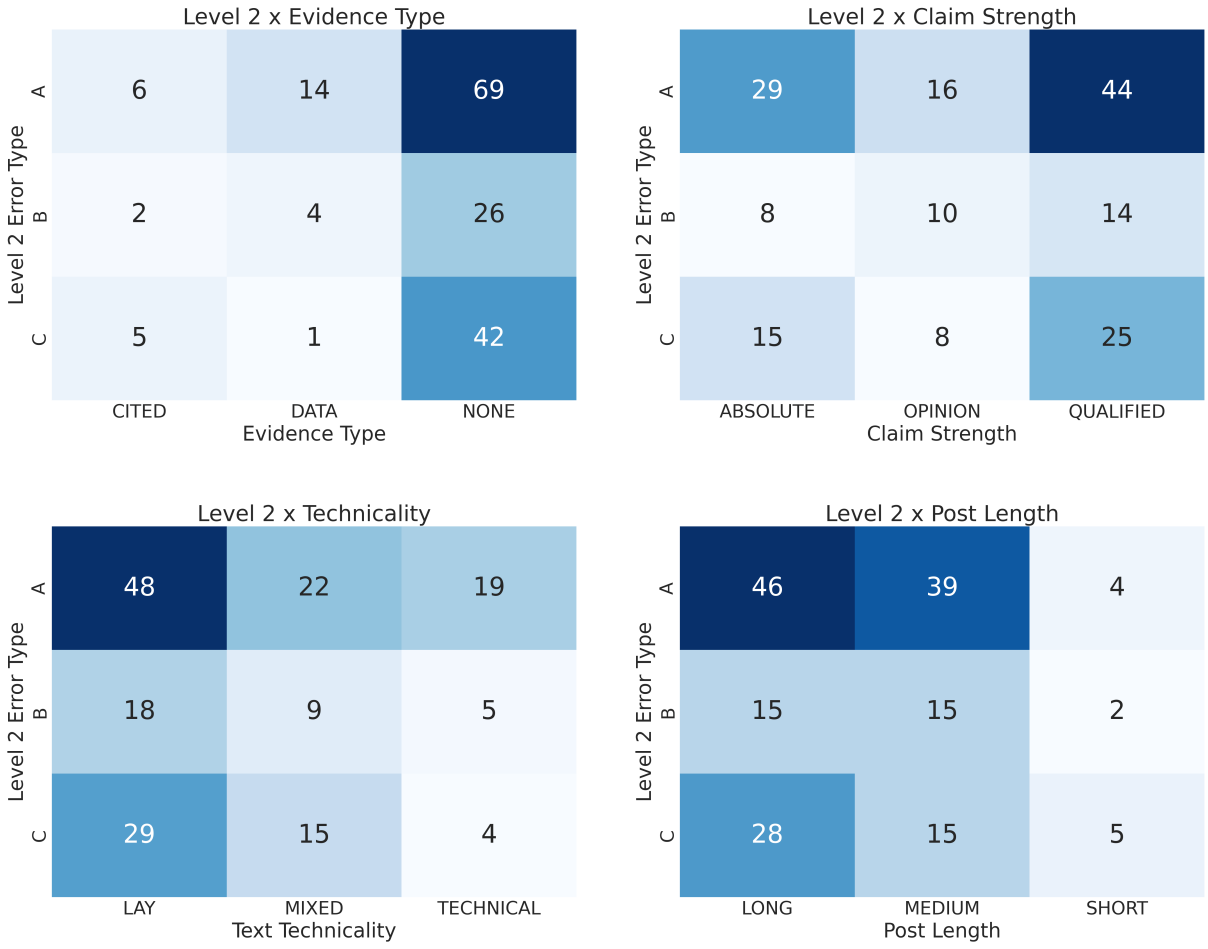

Figure 9: Heatmaps showing associations between Level 2 error mechanisms (content interpretation, reasoning, factual knowledge) and Level 3 post-level features (claim strength, evidence type, textual style, caption length) across misclassified predictions. Each panel represents a separate post-level feature.