TAIGR: Towards Modeling Influencer Content on Social Media via Structured, Pragmatic Inference

Anonymous ACL submission

Abstract

Health influencers play a growing role in shaping public beliefs[1], yet their content is often conveyed through conversational narratives and rhetorical strategies rather than explicit factual claims. As a result, claim-centric verification methods struggle to capture the pragmatic meaning of influencer discourse. In this paper, we propose **TAIGR** (**T**akeaway **A**rgumentation **I**nference with **G**rounded **R**eferences), a structured framework designed to analyze influencer discourse, which operates in 3 stages: (1) identifying the core influencer recommendation—*takeaway*; (2) constructing an argumentation graph that captures influencer justification for the *takeaway*; (3) performing factor graph-based probabilistic inference to validate the *takeaway*. We evaluate **TAIGR** on a content validation task over influencer video transcripts on health, showing that accurate validation requires modeling the discourse's pragmatic and argumentative structure rather than treating transcripts as flat collections of claims.

1 Introduction

Social media influencers increasingly act as informal knowledge brokers (Fabris et al., 2025), shaping public understanding and behavior in varied domains including health, finance, among others. Unlike traditional information sources, influencer-generated content is rarely presented as just factual reporting; instead, it is embedded in personal narratives, demonstrations and persuasive storytelling (Lee and Theokary, 2021a; Georgakopoulou and Garcés-Conejos Blitvich, 2024; Lim et al., 2025; Evans et al., 2026), often blurring the boundary between advice, belief and evidence. For instance, as shown in Figure 1, an objective claim like ("Vitamin D cures COVID") is often woven into a personal story ("When I tested positive...").

[1] https://hsph.harvard.edu/news/maha-shaping-public-health-messages-on-social-media/

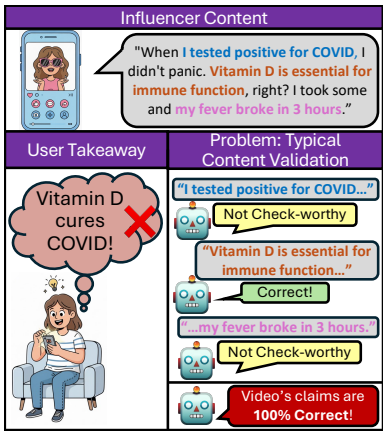

Figure 1: Treating influencer content as a series of claims fails to account for the pragmatic *Takeaway* conveyed to the user, and may lead to factually misleading results. We contend that un-"check-worthy" *anecdotes* may still be valuable to model, especially in their role of supporting an influencer's implicit takeaway.

Prior work has shown that such discourse can exert substantial influence on user attitudes and behavior, even when the users possess prior domain knowledge or skepticism (Cialdini, 2007; Ekinci et al., 2025; Evans et al., 2026). In the high-stakes health domain, this creates a need for developing robust systems that do not just assess surface-level correctness, but can also analyze how influencer messages are constructed and justified.

From a linguistic perspective, influencer content exposes a gap in the current NLP approaches. Existing literature has approached related problems through claim verification (Alhindi et al., 2018; Wadden et al., 2020; Kotonya and Toni, 2020; Li et al., 2025; Dmonte et al., 2025), argumentation mining (Mochales and Moens, 2011; Lawrence and Reed, 2019; Hidey et al., 2017) or social media modeling (Bozdag et al., 2025; Wang et al., 2019; Lee and Theokary, 2021b). However, treating these independently fails to provide a comprehensive view needed to understand the phenomenon of influencer discourse. An influencer's recommendation often manifests from anecdotes, credibility

cues ("my doctor said") or emotional framing; creating a gap between *what an influencer says and what audience is encouraged to believe or do*. For example, in Figure 1, the influencer never explicitly says "Vitamin D cures COVID", yet the user is encouraged to believe it. Our experiments reveal that this gap is where claim-centric verification pipelines break down, as exemplified in Figure 1.

Rather than treating discourse as a flat collection of verifiable propositions, we model it closer to how humans interpret and evaluate communicated information. The theory of epistemic vigilance (Sperber et al., 2010) posits that humans actively infer and evaluate communicated information through cognitive mechanisms that they use to protect themselves from being manipulated. These mechanisms operate along multiple dimensions: listeners infer *what is being communicated*, evaluate *the justification for the message*, and *assess the support reliability*.

Inspired by this perspective, we propose a structured framework, **TAIGR**: **T**akeaway **A**rgumentation **I**nference with **G**rounded **R**eferences, to model influencer discourse. First, **TAIGR** identifies the key *takeaway*, capturing **what** audiences internalize and act upon after watching influencer content. Then, it builds an argumentation structure, capturing **how** the influencer discourse supports the takeaway. Finally, augmenting this argumentation structure with external scientific evidence, it performs factor graph-based inference to **reason** over takeaway's trustworthiness.

To evaluate the efficacy of our approach in modeling influencer discourse, we adopt *content validation* as the downstream evaluation task. This task serves as a holistic signal, requiring models to fully understand and reason about an influencer's message. For this evaluation, we construct a dataset of 195 real-world TikTok videos, grounded in expert annotations from `ScienceFeedback`[2]. In addition, we conduct targeted human evaluation to assess the quality of intermediate components of the framework. On this dataset, **TAIGR** outperforms all baselines, achieving gains of up to $+9.7$ macro-F1 points, highlighting the importance of structured modeling for analyzing influencer discourse.

Beyond this evaluation, we apply **TAIGR** to 1,430 medical-domain TikTok videos and uncover two key patterns: (i) trustworthiness is largely decoupled from platform popularity, and (ii) trustworthiness is strongly shaped by rhetorical strategy, with premise-based reasoning associated with higher trust and anecdotal narratives with lower trust. These findings illustrate how structured modeling enables influencer discourse analysis at scale.

In summary, our contributions are:

- We propose **TAIGR**, a structured framework for analyzing influencer discourse, inspired by theories of epistemic vigilance.
- We demonstrate that **TAIGR** significantly outperforms baselines on the *content validation* task, achieving gains of up to $+9.7$ F1 points.
- We analyze influencer discourse on health at scale, showing that trustworthiness is independent of virality and is strongly influenced by rhetoric, revealing systematic differences between influencer and expert discourse.

2 Related Work

Prior work relevant to modeling influencer-generated content has largely focused on isolated aspects of the problem, rather than modeling influencer discourse holistically.

Claim Verification and Fact Checking. Early fact verification systems (Vlachos and Riedel, 2014; Volkova et al., 2017; Rashkin et al., 2017; Pérez-Rosas et al., 2018), primarily developed in the political domain, treat individual claims as the unit, without any external evidence. However, recent evidence-based methods either assume access to "gold evidence" (Alhindi et al., 2018; Atanasova et al., 2020) or rely on generic retrieval pipelines (Zheng et al., 2024; Russo et al., 2025). Similar approaches have been extended to scientific and health misinformation (Wadden et al., 2020; Pradeep et al., 2021; Kotonya and Toni, 2020; Vladika et al., 2024; Deka et al., 2023). More recently, LLM-based systems have been adapted to do claim decomposition, retrieval, and verification (Dmonte et al., 2025; Li et al., 2025; Wang et al., 2025). However, these approaches remain largely *claim-centric* and do not explicitly model the broader discourse structure through which influencers convey and justify health recommendations.

Argumentation Mining and Reasoning Structures. A complementary line of work focuses on recovering the structure of arguments (Mochales and Moens, 2011; Lawrence and Reed, 2019; Vecchi et al., 2021; Sviridova et al., 2025), and identifying claims and premises (Boltužić and Šnajder, 2016; Hidey et al., 2017; Dusmanu et al., 2017). More recently, studies (Zhu et al., 2025; Ng et al.,

[2] https://science.feedback.org

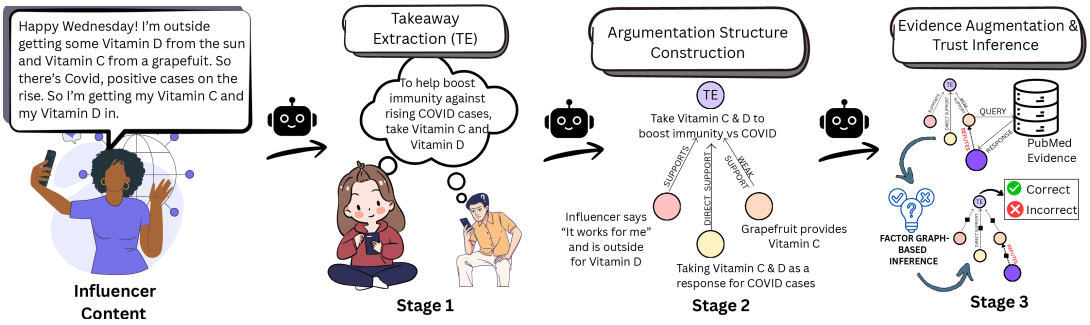

Figure 2: **TAIGR** overview: models influencer discourse in three stages: (1) *Takeaway* extraction, identifying the core recommendation; (2) *Argumentation Structure*, constructing a graph of supporting claims; and (3) *Trust Inference*, validating the recommendation using external evidence.

2025) leverage LLMs to construct and reason over argument graphs for claim verification. While these methods capture *how* claims are supported, they are not developed to address the noisy, implicit, and persuasion-driven nature of influencer discourse.

Social Media Modeling. Prior work examined how persuasion (Wang et al., 2019; Lee and Theokary, 2021b; Akhynko et al., 2025; Bozdag et al., 2025), credibility (Liu and Zheng, 2024), and influence (Sharma et al., 2022; Fang and Wang, 2022) manifest in social media, offering valuable insights into rhetorical strategies.

In contrast to studies centered on social network analysis (Del Tredici et al., 2019; Razis et al., 2020), our work focuses on the *linguistic and argumentative structure* of influencer-generated content itself. By integrating takeaway identification, argumentation modeling, and evidence-grounded reasoning, we bridge insights from fact checking, argument mining, and social media analysis to analyze influencer discourse in an end-to-end manner.

3 TAIGR

Drawing inspiration from the theory of epistemic vigilance (Sperber et al., 2010), which characterizes how humans evaluate communicated information, we propose a structured framework, **TAIGR**, to analyze influencer discourse. It has three interdependent stages, examining different aspects of the discourse: (1) Takeaway: Influencer's key recommendation; (2) *Argumentation Structure*: Represents the influencer's justification for the *takeaway* using an argumentation graph; (3) *Trust Inference*: Infers the degree to which the *takeaway* is corroborated by external evidence;

Figure 2 shows an overview of **TAIGR** and described in detail below. It operates over the text transcriptions of influencer videos, obtained using OpenAI's Whisper API[3].

Content	Type
Transcript-1: [At urgent care] I was diagnosed with COVID. [was sent home..called my nursing friends and doctor] They both recommended the ivermectin [took 1st dose..hoping] it relieves my symptoms soon. **Takeaway:** If diagnosed with COVID, consider taking ivermectin as a potential treatment to relieve symptoms.	Implicit
Transcript-2: Go get your HPV vaccine. Now, go get your [children], age 11 to 12. This vaccine has been proven to prevent [...] HPV related cancer, cervical cancer. **Takeaway:** Go get your HPV vaccine and ensure that your children aged 11 to 12 also receive it, as it has been proven to prevent HPV-related cancers such as cervical cancer.	Explicit

Table 1: Example for *Implicit* vs. *Explicit* takeaway type.

3.1 Takeaway

We define *Takeaway* as the influencer's core actionable recommendation. It represents the central message the audience is intended to internalize and act upon, such as adopting a health routine, purchasing a product, or following a specific protocol.

The identification of *takeaway* provides a semantic anchor for analyzing how the influencer supports or justifies the conveyed message. Similar to (Alhindi et al., 2018), who model claim's *justification* of a claim to improve downstream performance, we argue that accurately identifying the takeaway is a necessary prerequisite for effectively modeling influencer discourse.

Takeaways may be expressed either *explicitly* or *implicitly* (Table 1). Explicit takeaways are directly stated as advice or recommendations (e.g., 'Get your vaccine'). Implicit takeaways, by contrast, are inferred from demonstrations, personal experiences, or causal narratives. In such cases, there is a gap between what the influencer says and what they imply. Despite their indirect form, they can be equally influential and persuasive.

[3] https://pypi.org/project/openai-whisper/

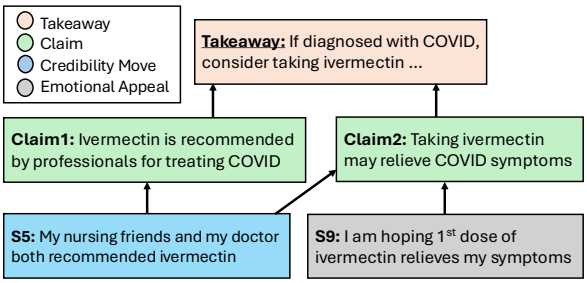

Figure 3: Argumentation graph for *transcript-1* (Table 1), with node types and directed support edges.

Our takeaway extraction component captures both cases. Given the transcript of a video, we prompt an LLM to identify the key *takeaway* from the transcript and classify whether it is *explicit* or *implicit*. Due to the short nature of the videos, we assume that each video contains exactly one *takeaway*. Prompts used for Takeaway identification and classification are in Appendix E.1.

3.2 Argumentation Structure

Having identified the *takeaway*, we develop a representation to analyze how it is supported within the transcript by recovering its argumentative structure, linking claims to the central recommendation.

Following prior work in argumentation mining and claim–premise modeling (Mochales and Moens, 2011; Hidey et al., 2017; Sviridova et al., 2025), we represent each transcript as a directed argumentation graph. We segment the transcript into standalone statements, assign rhetorical roles, identify claims in the transcript, and build the structure by inferring the relationship between them.

Notation. Let a transcript be denoted by a set of statements $\mathcal{S} = \{s_1, \ldots, s_n\}$ and a set of claims $\mathcal{C} = \{c_1, \ldots, c_m\}$. The extracted takeaway is denoted as T. Our objective is to construct a directed graph $G = (V, E)$, where $V = \mathcal{S} \cup \mathcal{C} \cup \{T\}$, and edges represent supportive argumentative relations.

Preprocessing. Given a transcript, we first segment it into self-contained *statements* by resolving coreferences, using an LLM. These statements serve as the atomic units of influencer discourse.

Each statement $s_i \in \mathcal{S}$ is then classified by its rhetorical function, reflecting common persuasive strategies used by influencers. Following prior work (Hidey et al., 2017), we label statements as one of: *premise* (logos; appeals to reasoning), *anecdotal evidence*, *credibility move* (ethos), *emotional appeal* (pathos), or *None*. These labels are assigned using an LLM prompted to identify the argumentative role of each statement given the transcript. Figure 3 shows an example of a statement $S5$ classified as *credibility move*.

To capture both explicit and implicit reasoning (Sviridova et al., 2025), we further extract a set of *claims* $\mathcal{C}$, checkworthy propositions in the transcript, using an LLM. For each claim c_j, the LLM also identifies the subset of statements $\mathcal{S}_{c_j} \subseteq \mathcal{S}$ that support or give rise to the claim, yielding a local argumentative structure in which claims are justified by rhetorically typed statements. For instance, $Claim1$ is directly derived from $S5$ in Figure 3.

Building the argumentation graph. We represent the influencer's reasoning as a directed argumentation graph rooted at *takeaway* node T. Nodes correspond to the extracted argumentative units (claims, statements, takeaway), and directed edges represent inferred support relations between them.

We begin by identifying candidate edges from each claim $c_j \in \mathcal{C}$ to the takeaway T. For each pair (c_j, T), we prompt an LLM to classify the relationship as *direct support*, *weak support*, or *no support*. Following prior work showing LLMs can meaningfully express uncertainty (Kadavath et al., 2022), we compute a continuous edge weight using the model's class probabilities: $w(c_j, T) = 1.0 * P(\text{direct support}) + 0.5 * P(\text{weak support})$. Edges whose scores exceed a fixed threshold are added to the graph, yielding an initial set of claim-to-takeaway support links.

We then iteratively expand the graph to capture deeper argumentative structure. For each node already connected to the graph, we evaluate whether remaining claims provide support to it, using the same classification and weighting procedure. Edges that meet the threshold are added, allowing claims to support intermediate claims as well as the takeaway.

Finally, we attach remaining statement-level nodes using the same process, continuing expansion until all candidate nodes have been evaluated.

The resulting graph captures the inferred structure of how an influencer's claims collectively support the takeaway. Figure 3 illustrates an example argumentation graph for *transcript-1* (Table 1).

Implementation details, thresholds, and prompts are provided in Appendix B.2 and Appendix E.1.

3.3 Trust Inference

After constructing the argumentation graph, we estimate whether the *takeaway* should be trusted by assessing the factual accuracy of the arguments

used to support it. Rather than modeling trust as an intrinsic property of the influencer, we operationalize it as the degree to which the supporting arguments align with external scientific evidence. Under this formulation, a takeaway is considered more trustworthy when its argumentative support is consistent with reliable biomedical literature.

We validate using two steps. First, we retrieve and classify scientific evidence from PubMed[4] to augment the argumentation graph. Second, we reason over this evidence-augmented graph to infer the trustowrthiness of the *takeaway*.

3.3.1 Evidence Retrieval and Classification

Our evidence retrieval pipeline follows (Li et al., 2025) and consists of four steps: identifying checkworthy nodes, expanding search queries, retrieving and reranking evidence, and classifying evidence; described below in detail. Implementation details and prompts are provided in Appendix B.2 and E.1.

Checkworthy Node Identification. Not all nodes in the argumentation graph are verifiable. We restrict evidence retrieval to *Claim* and *Premise* nodes, excluding rhetorical categories such as anecdotal evidence or emotional appeals. For each candidate node, an LLM determines whether it is *PubMed-checkworthy*, i.e., whether it expresses a proposition that can plausibly be supported or refuted by biomedical literature.

Query Expansion. To improve recall and capture both supporting and opposing evidence (Ousidhoum et al., 2022), we expand each checkworthy node into multiple queries. Specifically, an LLM generates up to five supporting and five opposing queries per node, encouraging retrieval of evidence with diverse scientific stances.

Evidence Retrieval and Reranking. For each expanded query, we retrieve the top-K (K=100) candidate PubMed documents using `ColBERT-v2` (Santhanam et al., 2022). To filter weakly relevant results, we apply a `monoT5` reranker (Nogueira et al., 2020) (`monot5-base-med-msmarco`) and retain top-30 evidence documents per node for downstream classification.

Evidence Classification. Each retrieved PubMed article is classified by an LLM into one of five categories: {*strong support, weak support, neutral, weak oppose, strong oppose*}. Along with the class label, we additionally compute a continuous edge weight using the LLM's class probabilities: $e = -1.0 \cdot P(\text{strong oppose}) - 0.5 \cdot P(\text{weak oppose}) + 0 \cdot P(\text{neutral}) + 0.5 \cdot P(\text{weak support}) + 1.0 \cdot P(\text{strong support})$. This formulation jointly captures the polarity and strength of the evidence, yielding a score in $[-1, 1]$ that quantifies how strongly an evidence node supports or contradicts the associated claim or premise.

The resulting evidence nodes and weighted support or attack edges are added to the argumentation graph, producing an evidence-augmented structure, used for inferring the takeaway's trustworthiness.

3.3.2 Factor graph based content validation

Given the evidence-augmented argumentation graph, we infer how much the influencer's *takeaway* should be trusted by propagating evidence credibility through the argumentative structure. We treat PubMed evidence as the most reliable source and use it as an anchor from which trust propagates to claims and ultimately to the takeaway.

To enforce global consistency over this structure, we formulate trust inference as a factor graph and perform approximate inference using AD3 (Martins et al., 2015). This formulation allows us to jointly reason about local evidence strength and global argumentative relations, while explicitly modeling both support and contradiction.

Variables and states. Each node v in the augmented argumentation graph corresponds to a discrete random variable Y_v, representing its trustworthiness. Trust is discretized into 11 states $\{0, 0.1, \ldots, 1.0\}$. PubMed evidence nodes are assigned a hard constraint $Y_v = 1.0$, reflecting their assumed high credibility.

Inference objective. We seek a trust assignment that maximizes the sum of unary and binary potentials:

$$\arg\max_{y_1,\ldots,y_m} \sum_{v \in V} \theta_v(y_v) + \sum_{(u,v) \in E} \theta_{uv}(y_u, y_v),$$

where unary potentials encode prior trust assumptions and binary potentials specify how trust values of connected nodes should influence one another.

Unary potentials. Unary potentials enforce hard constraints for PubMed evidence nodes and assign lower prior trust to rhetorical or non-factual nodes. All remaining nodes receive uniform unary potentials, allowing their trust to be determined by surrounding evidence and structure.

Binary potentials. Binary potentials define how trust flows between connected nodes in the argumentation graph, with stronger support or opposi-

[4] https://pubmed.ncbi.nlm.nih.gov/

tion relations leading to greater influence on the final trust assignment.

For a **support** edge, we define: $\theta^{\text{support}}(y_u, y_v) = \beta \cdot w \cdot \big(1 - (y_u - y_v)^2\big)$, which encourages similar trust assignments between supporting nodes. As a result, trust from credible evidence propagates through supporting claims towards the takeaway.

For an **attack** edge, we define: $\theta^{\text{attack}}(y_u, y_v) = \gamma \cdot |w| \cdot \big(1 - (y_u + y_v - 1)^2\big)$, which encourages complementary trust assignments. This formulation ensures strong trust in one node implies reduced trust in the opposing node, preventing contradictory evidence from jointly receiving high confidence.

Here, β and γ control the relative influence of support and attack relations[5].

Inference and validation. We solve the resulting factor graph using AD3, which relaxes the discrete problem into a linear program and efficiently produces approximate marginal trust assignments. The inferred trust value of the *takeaway* node serves as our content validity score: values in $[0, 0.5]$ are labeled *incorrect*, $[0.5, 0.7]$ as *partially correct*, and > 0.7 as *correct*. Implementation details and hyperparameters are provided in Appendix B.2.

4 Data

Following (Shang et al., 2025a), we construct a dataset of health-related TikTok videos using hashtag-based search. The hashtags are derived from `ScienceFeedback` website, ensuring we primarily capture health-related influencer discourse (see Appendix B.1 for details). To focus on influencer-generated content with meaningful audience reach, we retain only videos posted by creators with at least 1,000 followers. The resulting dataset consists of 1,430 videos. Table 2 summarizes the distribution of videos by influencer popularity and by takeaway type. We note that a substantial portion of the dataset contains *implicit* takeaways, reflecting the indirect nature of health advice commonly observed in influencer content.

5 Evaluation

We evaluate our framework using human annotations and an automated *content validation* task, which requires a model to judge whether an influencer's content is {*correct, partially correct, incorrect*} given the video transcript.

Crucially, solving this task requires more than surface-level pattern matching: a model must recover the influencer's intended recommendation, understand how it is argued for, and assess whether the supporting claims are scientifically valid. We therefore use content validation task as an end-to-end evaluation signal for the proposed framework. We use `GPT-4.1-mini`[6] for all LLM calls and report results from a single execution of the model.

By Influencer Popularity			
Category	**#Videos**	**Category**	**#Videos**
0 – 50K	603	250K – 500K	155
50K – 100K	218	500K – 1M	112
100K – 250K	220	> 1M	122
By Takeaway Type			
Implicit	533		
Explicit	897		
Total Dataset	**1,430**		

Table 2: Dataset Distribution.

5.1 Experimental Setup

Annotated Dataset. To support content validation, we construct an annotated dataset by aligning influencer videos with expert-reviewed medical claims. We use `ScienceFeedback`, which contains real-world claims verified by medical experts and labeled into 18 fine-grained veracity categories.

For each video transcript (Section 4), we extract its central *takeaway* and identify semantically corresponding claims from `ScienceFeedback`. We retain pairs where the expert-annotated claim entails the extracted takeaway, yielding 185 candidates. After manual verification to ensure alignment semantic equivalence, we obtain 166 high-quality {*takeaway*, claim} pairs.

Due to limited data and label skewness, we collapse the 18 original categories into three coarse-grained classes: {*correct, partially correct, incorrect*}, and assign each video the label of its associated takeaway. To increase coverage of high-confidence correct cases, we further augment the dataset with 29 videos from the `Cleveland Clinic`, a medically authoritative source, all labeled as *correct*. In total, the evaluation set contains 195 labeled video transcripts. Additional details are provided in Appendix C.1.

Baselines. We compare our framework against the following baselines (detailed in Appendix C.2), which differ in how explicitly they model influencer content and its underlying reasoning.

Claim-centric RAG. This baseline decomposes a transcript into factual claims using an LLM, re-

[5] We set $\beta = 0.3, \gamma = 5$.

[6] https://platform.openai.com/docs/models/gpt-4.1-mini

Method	F1 (Macro)	Class-wise F1			F1 by Takeaway Type	
		Incorrect	Partial	Correct	Implicit	Explicit
Claim-centric RAG	0.3	0.05	0.32	0.53	0.21	0.31
RAG-w-Takeaway	0.42	0.2	0.39	0.69	0.44	0.41
LOKI (Li et al., 2025)	0.43	0.71	0.29	0.29	0.35	0.43
TAIGR (No Structural Propagation)	0.46	0.66	0.23	0.47	0.47	0.43
TAIGR	**0.52**	0.72	0.42	0.41	**0.65**	**0.49**

Table 3: Content validation task results. **(a) Main Results:** The first columns show the overall Macro-F1 and the class-wise performance breakdown (Incorrect/Partial/Correct). We assessed statistical significance using a paired bootstrap test (10000 iterations). Testing the hypothesis that our model outperforms LOKI, we achieved a statistically significant result ($p = 0.039 < 0.05$). **(b) Takeaway Analysis:** The rightmost columns compare performance across takeaway type.

trieves PubMed evidence for each claim, and classifies the retrieved evidence as supporting or opposing. The LLM then predicts the overall veracity of the transcript based on the aggregated evidence. This setup does not explicitly model the central recommendation conveyed by the influencer.

RAG-w-Takeaway. This baseline introduces limited structure by first extracting the *takeaway* and retrieving PubMed evidence for it. The retrieved evidence is classified as supporting or opposing, and the LLM is prompted to predict the veracity of the *takeaway* given this evidence. This baseline isolates the impact of takeaway-aware retrieval without modeling claim interactions.

LOKI. We adapt LOKI (Li et al., 2025), a claim-centric verification system, to our setting. LOKI decomposes transcripts into claims, filters uncheckworthy ones, expands retrieval queries to improve recall, and verifies each claim independently using PubMed evidence. We follow (Li et al., 2025) for threshold selection for the *incorrect* class and tune the *partially correct* threshold (Appendix C.2).

Metrics. We evaluate all models using a standard classification metric: macro-averaged F1-score, to account for class imbalance.

5.2 Results and Ablation

Table 3 presents results on the content validation task. Overall, our framework achieves the highest macro F1-score, outperforming all competing baselines by a margin of up to +9.7 points.

Main Results. Among the baselines, **Claim-centric RAG** performs poorly, indicating that treating influencer transcripts as collections of loosely related factual claims is insufficient for validating their overall message. Introducing minimal structure via **RAG-w-Takeaway** substantially improves performance, highlighting the importance of explicitly modeling what the influencer is recommending.

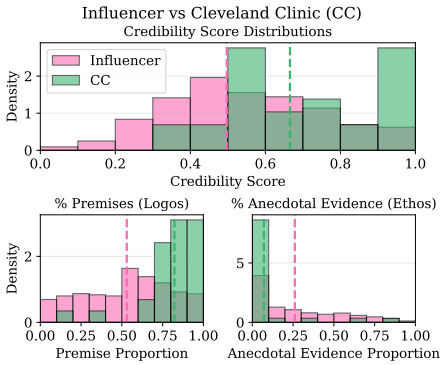

Figure 4: Rhetorical distributions of TikTok influencers vs. experts: expert content relies on premises and achieves higher credibility, while influencer content emphasizes anecdotes.

LOKI performs comparably to **RAG-w-Takeaway**, despite using a more sophisticated claim-filtering and query-generation pipeline. However, both baselines fall short of our method. While LOKI verifies claims individually and aggregates them uniformly, it does not model their interdependencies or relative influence on the final takeaway. **TAIGR** explicitly captures these dependencies, enabling more accurate trust inference. The improvement over the strongest baseline (LOKI) is statistically significant under a paired bootstrap test with 10,000 iterations ($p = 0.039 < 0.05$).

Impact of Takeaway Type. Table 3(b) shows performance by takeaway type. Our method outperforms all baselines for both *explicit* and *implicit* takeaways, with particularly large gains for implicit cases. For implicit takeaways, the improvement over LOKI is statistically significant (Appendix D).

This gap is primarily attributable to LOKI's reliance on surface-level claim decomposition. Influencer transcripts are conversational and often imply recommendations indirectly. When the takeaway is implicit, LOKI frequently fails to recover a claim that corresponds to the intended recommendation, leading to ineffective evidence retrieval. Table 4

illustrates this failure mode: for *transcript-1*, LOKI extracts claims centered on peripheral details (e.g., who recommended ivermectin), but fails to capture the implied recommendation of using ivermectin to treat COVID-19. Consequently, relevant evidence is not retrieved, resulting in incorrect validation.

Impact of Argument Structure Modeling. We evaluate the effect of modeling interactions between claims by comparing ***TAIGR*** *(No Structural Propagation)*, which performs independent claim validation without graph-based inference, against our full method. As shown in Table 3(a), this variant outperforms LOKI, reflecting improved claim extraction and evidence retrieval. However, it remains substantially worse than the full model (-6 F1 points), demonstrating the importance of modeling claim dependencies and trust propagation through the argumentation graph.

Taken together, these results show that performance gains arise not from any single component, but from the integration of (i) takeaway-based modeling, (ii) argumentation structure, and (iii) structured trust propagation over retrieved evidence.

Human Validation. As a qualitative sanity check of our framework's intermediate components, we conduct human validation on a randomly sampled set of 100 videos. We evaluate the LLM's ability to (i) correctly identify the video's *takeaway*, (ii) determine whether the takeaway is expressed *explicitly* or *implicitly*, and (iii) assign rhetorical function labels to individual statements. Human judgments show that the LLM identifies the takeaway with 75% accuracy, classifies implicit versus explicit takeaways with 79% accuracy, and assigns rhetorical labels with 72% accuracy. Despite the noisy and conversational nature of influencer discourse, these results indicate that LLMs can reliably recover the semantic intent and rhetorical structure required by our framework.

6 Analysis of Rhetoric and Credibility

Having established that our trust score reliably reflects content veracity, we apply **TAIGR** to all 1,430 influencer videos to analyze how influencer popularity, rhetorical strategy, and credibility interact. Throughout this section, we use *credibility* as a shorthand to denote the *trustworthiness* score assigned by our model.

Popularity $\neq$ Credibility. Figure 5 reports correlations between video features and credibility. We observe a **near-zero correlation between our credibility score and engagement metrics**$-$# Followers ($r = 0.05$), # Plays ($r = 0.07$), indicating that neither account popularity nor video virality is predictive of medical claim validity. In other words, evidence-backed health advice is no more likely to be promoted or consumed than low-credibility content, suggesting that platform dynamics are largely agnostic to veracity.

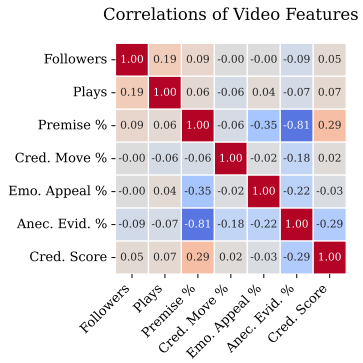

Figure 5: Is viral content more credible? We plot the correlation matrix of video features, showing that engagement metrics are decoupled from credibility (near-0 correlation).

Anecdotes Reduce Credibility. Rhetorical structure is a significant predictor of a video's trustworthiness. We find that Credibility Score correlates positively with the use of Premises ($r = 0.29$) and negatively with Anecdotal Evidence (-0.29). This suggests that videos grounded in logical claims are inherently more credible than those relying on personal testimonials. While these techniques are often used in opposition ($r = -0.81$), their individual impact on veracity is also opposite.

Influencers vs. Experts. To contextualize our findings, we compare the influencer videos against the "gold standard" Cleveland Clinic (CC) videos. Figure 4 illustrates the rhetorical divide. The expert content (CC) is dominated by premises (Logos), resulting in a high-credibility distribution. Conversely, the influencer distribution is shifted heavily toward anecdotal evidence (Ethos), confirming that reliance on personal testimonials over structural arguments drives the credibility gap between the two, per our model. Appendix A.3 shows further analysis by specific health topics.

7 Conclusion

We introduce **TAIGR**, a structured framework for validating influencer content by modeling the *takeaway*, its supporting arguments, and external evidence grounding. TAIGR outperforms strong baselines, particularly on videos with implicit recommendations, while enabling interpretable analysis.

Ethics Statement

This work studies health information in influencer discourse using transcripts derived from publicly available TikTok videos. We focus on public accounts and collect from influencers with an average $\sim$345,000 followers per influencer, aiming to analyze high-reach public-facing content rather than private individuals. We do not access private accounts, bypass access controls, or collect deleted content.

Privacy and identifiability. To minimize privacy risks associated with multimedia data, our modeling and analysis operate only on textual transcripts (generated via Whisper) and do not use or release raw video, audio, images, or biometric signals (e.g., faces or voices). We avoid presenting identifying details about individual creators and report findings in aggregate.

Data use, redistribution, and platform policies. Platform terms may restrict automated collection and redistribution of content. We therefore do not publicly release raw TikTok content or direct copies of transcripts. To support reproducibility, we will provide access to derived textual data to researchers upon request and agreement not to redistribute.

Potential harms and misuse. Studying misinformation may inadvertently amplify harmful claims or enable adversarial use. We mitigate this by (i) focusing on analysis and validation rather than promoting specific creators or claims, (ii) avoiding the release of content that would facilitate re-sharing of misinformation, and (iii) grounding validation in reputable biomedical literature (PubMed). We believe the public-health benefit of understanding and detecting persuasive health misinformation outweighs the residual risks under these safeguards.

Medical disclaimer. This work is intended for research; it does not provide medical advice, and model outputs should not be used as clinical guidance.

Limitations

Retrieved evidence. Our framework retrieves external evidence exclusively from PubMed. This design choice prioritizes source reliability and consistency in the health domain, but it also limits coverage. Restricting retrieval to a single curated source may exclude relevant evidence available in other scientific repositories or authoritative guidelines. Expanding retrieval to the open web introduces additional challenges, including variability in source trustworthiness and the need to model it, as well as dependence on evidence curated by human fact-checkers (Chen et al., 2024). Addressing these trade-offs remains an important direction for future work.

Text-only modality. Our analysis operates solely on textual transcriptions of influencer videos. While this enables understanding influencer content to some degree, it ignores potentially informative cues from other modalities, such as visual demonstrations, on-screen text, tone, and gestures. Prior work (Shang et al., 2025b) has shown that multimodal signals can play an important role in interpreting influencer discourse, especially in health misinformation. Incorporating multimodal reasoning remains an important direction for extending the framework.

Modular design. Our framework is intentionally modular, enabling fine-grained inspection of intermediate representations such as takeaways and argumentation structures. While this design introduces the possibility of error propagation across stages, it also allows errors to be explicitly identified, analyzed, and addressed. We address this issue through human evaluation of intermediate components of our framework.

Dataset limitation. There is a lack of publicly available, large-scale datasets specifically designed to analyze influencer discourse and its pragmatic structure. As a result, we construct a proxy dataset aligned with `ScienceFeedback`, which provides expert-annotated assessments of scientific validity. While this enables systematic evaluation, it may not fully capture the diversity of influencer styles, platforms or non-health domains encountered in real-world settings.

Modeling Assumption. Motivated by the short-form nature of most influencer videos, we assume that each video consists of only one dominant takeaway. However, longer or more complex content may convey multiple recommendations or nuanced messages. Extending the framework to jointly model multiple takeaways within a single video is an important direction for future work.

8 Statement of AI Assistant Usage

The development of the codebase was partially supported by an AI assistant for debugging, and the

manuscript was refined with an AI assistant.

References

Kateryna Akhynko, Oleksandr Kosovan, and Mykola Trokhymovych. 2025. Hidden persuasion: Detecting manipulative narratives on social media during the 2022 Russian invasion of Ukraine. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 194–202, Vienna, Austria (online). Association for Computational Linguistics.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Filip Boltužić and Jan Šnajder. 2016. Fill the gap! analyzing implicit premises between claims from online debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 124–133, Berlin, Germany. Association for Computational Linguistics.

Nimet Beyza Bozdag, Shuhaib Mehri, Xiaocheng Yang, Hyeonjeong Ha, Zirui Cheng, Esin Durmus, Jiaxuan You, Heng Ji, Gokhan Tur, and Dilek Hakkani-Tür. 2025. Must read: A systematic survey of computational persuasion. *Preprint*, arXiv:2505.07775.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587, Mexico City, Mexico. Association for Computational Linguistics.

Robert B Cialdini. 2007. *Influence: The psychology of persuasion*, volume 55. Collins New York.

Pritam Deka, Anna Jurek-Loughrey, and Deepak P. 2023. Multiple evidence combination for fact-checking of health-related information. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 237–247, Toronto, Canada. Association for Computational Linguistics.

Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You Shall Know a User by the Company It Keeps: Dynamic Representations for Social Media Users in NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4707–4717, Hong Kong, China. Association for Computational Linguistics.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2025. Claim Verification in the Age of Large Language Models: A Survey. *arXiv preprint*. ArXiv:2408.14317 [cs].

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

Yuksel Ekinci, Shubhankar Dam, and Georgia Buckle. 2025. The dark side of social media influencers: A research agenda for analysing deceptive practices and regulatory challenges. *Psychology & Marketing*, 42(4):1201–1214.

Sophie Evans, Kelly Lambert, Adrian Dinale, Myah Quinn, and Denelle Cosier. 2026. Unqualified advice and product promotions: Analysis of health and nutrition content on social media consumed by young adults. *Nutrients*.

Stephanie Russo Fabris, Jose Adão Carvalho Nascimento Júnior, Mairim Russo Serafini, and Luis Felipe Dias Lopes. 2025. Self-care or misinformation? investigating the herbal wellness movement on tiktok.

Xing Fang and Tianfu Wang. 2022. Using Natural Language Processing to Identify Effective Influencers. *International Journal of Market Research*, 64(5):611–629. Publisher: SAGE Publications.

Alexandra Georgakopoulou and Pilar Garcés-Conejos Blitvich. 2024. Influencer discourse: Affective relations and identities.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 7740–7754, Online. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818. Place: Cambridge, MA Publisher: MIT Press.

Michael T Lee and Carol Theokary. 2021a. The superstar social media influencer: Exploiting linguistic style and emotional contagion over content? *Journal of Business Research*, 132:860–871.

Michael T. Lee and Carol Theokary. 2021b. The superstar social media influencer: Exploiting linguistic style and emotional contagion over content? *Journal of Business Research*, 132:860–871.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2025. Loki: An open-source tool for fact verification. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 28–36, Abu Dhabi, UAE. Association for Computational Linguistics.

Weng Marc Lim, Kian Yeik Koay, and Chee Wei Cheah. 2025. Commercialization of influencer content on social media: the roles of parasocial relationships and sponsorship disclosures. *Journal of Strategic Marketing*, 33(5):658–684.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu and Xiaoyong Zheng. 2024. The persuasive power of social media influencers in brand credibility and purchase intention. *Humanities and Social Sciences Communications*, 11(1):1–12.

André FT Martins, Mário AT Figueiredo, Pedro MQ Aguiar, Noah A Smith, and Eric P Xing. 2015. Ad3: Alternating directions dual decomposition for map inference in graphical models. *The Journal of Machine Learning Research*, 16(1):495–545.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial intelligence and law*, 19(1):1–22.

Ming Pok Ng, Junqi Jiang, Gabriel Freedman, Antonio Rago, and Francesca Toni. 2025. Marge: Meshing argumentative evidence from multiple large language models for justifiable claim verification. *Preprint*, arXiv:2508.02584.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Scientific claim verification with VerT5erini. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 94–103, online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Gerasimos Razis, Ioannis Anagnostopoulos, and Sherali Zeadally. 2020. Modeling Influence with Semantics in Social Networks: A Survey. *ACM Comput. Surv.*, 53(1):7:1–7:38.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Daniel Russo, Stefano Menini, Jacopo Staiano, and Marco Guerini. 2025. Face the facts! evaluating rag-based pipelines for professional fact-checking. *Preprint*, arXiv:2412.15189.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Lanyu Shang, Yang Zhang, Yawen Deng, and Dong Wang. 2025a. Multitec: A data-driven multimodal short video detection framework for healthcare misinformation on tiktok. *IEEE Transactions on Big Data*, 11(5):2471–2488.

Lanyu Shang, Yang Zhang, Yawen Deng, and Dong Wang. 2025b. MultiTec: A Data-Driven Multimodal Short Video Detection Framework for Healthcare Misinformation on TikTok. *IEEE Transactions on Big Data*, 11(5):2471–2488.

Amit Sharma, Neha Patel, and Rajesh Gupta. 2022. Leveraging Machine Learning Algorithms and Natural Language Processing for Enhanced AI-Driven Influencer Campaign Analytics. *European Advanced AI Journal*, 11(10).

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Preprint*, arXiv:2004.09297.

Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. 2010. Epistemic vigilance. *Mind & language*, 25(4):359–393.

Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2025. Mining implicit arguments for reasoning: A survey. *Argument & Computation*, page 19462174251344764. Publisher: SAGE Publications.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards Argument Mining for Social Good: A Survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1338–1352, Online. Association for Computational Linguistics.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi N. Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2025. OpenFactCheck: Building, benchmarking customized fact-checking systems and evaluating the factuality of claims and LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11399–11421, Abu Dhabi, UAE. Association for Computational Linguistics.

Liwen Zheng, Chaozhuo Li, Xi Zhang, Yu-Ming Shang, Feiran Huang, and Haoran Jia. 2024. Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9274–9281, Bangkok, Thailand. Association for Computational Linguistics.

Yuqicheng Zhu, Nico Potyka, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Dongzhuoran Zhou, Evgeny Kharlamov, and Steffen Staab. 2025. Argrag: Explainable retrieval augmented generation using quantitative bipolar argumentation. *Preprint*, arXiv:2508.20131.

A Extended Results

A.1 Impact of Takeaway Type

Table 4 shows the comparison between LOKI, a claim-centric verifier, against our structured approach towards content validation of influencer discourse. In the implicit cases, we observe that LOKI typically fails to identify the *takeaway*; in this case, fails to associate COVID-19 with ivermectin, and this in turn fails to retrieve relevant evidence, resulting in incorrect classification.

A.2 Human Validation Setup

We conduct a human validation study on a randomly sampled set of 100 videos to assess the accuracy of LLM predictions in: (i) identifying the correct *takeaway*, (ii) determining whether the takeaway is expressed *explicitly* or *implicitly*, and (iii) assigning rhetorical function labels to individual statements in the transcript.

Baseline (LOKI)	TAIGR
Decomposed Claim: The nursing friends and doctor recommended ivermectin. **Generated Retrieval Queries:** • What are the clinical indications for ivermectin use according to peer-reviewed guidelines? • Is there evidence that supports recommendations by nursing staff?	**Extracted Claim:** Ivermectin is recommended by professionals for treating COVID symptoms. **Generated Retrieval Queries:** • Clinical trials of ivermectin prescribed by doctors for COVID-19 symptoms. • Adverse effects reported from ivermectin use in COVID-19 patients.

Table 4: Comparison of LOKI (baseline) against **TAIGR** for the *transcript-1* (Table 1). None of the LOKI-decomposed claims account for the implication of *treating COVID-19* with ivermectin. Consequently, it fails to correctly classify the transcript when the takeaway type is *implicit*.

Annotations were collected via a Qualtrics survey. Annotators were first provided with a definition of a *takeaway*, followed by the transcript of a video and the LLM-predicted takeaway. They were asked: *"Is this a valid takeaway from the video?"* (Yes/No). When the takeaway was judged valid, annotators were subsequently asked: *"Is the predicted takeaway type (explicit or implicit) consistent with the video?"* (Yes/No).

Independently of takeaway evaluation, annotators were also shown a random subset of $3-5$ statements from each transcript, together with their predicted rhetorical function labels, and were asked to assess whether the labels were appropriate.

Across the annotated sample, the LLM correctly identified the takeaway in 75% of cases, classified explicit versus implicit takeaways with 79% accuracy, and assigned rhetorical function labels with 72% accuracy.

In this case, human annotators were graduate-STEM students under the age of 40, who were not the authors of the paper.

A.3 Analysis By Health Topic

In Figure 6, we show all of our framework's average scores for the full set of videos, separated by the health topic. The health topic is determined by the original hashtag(s) used to obtain each video.

We see that for credibility score itself, topics like General Health and Preventative Care are seen as more credible. We hypothesize that this is because these are topics that real medical professionals are willing to provide advice for over social media videos. We see that topics of Medication scores much lower, indicating that many influencers may be making dubious drug recommendations. We see this especially in the Anecdotal Evidence plot, where Medication has a very high proportion of anecdotes.

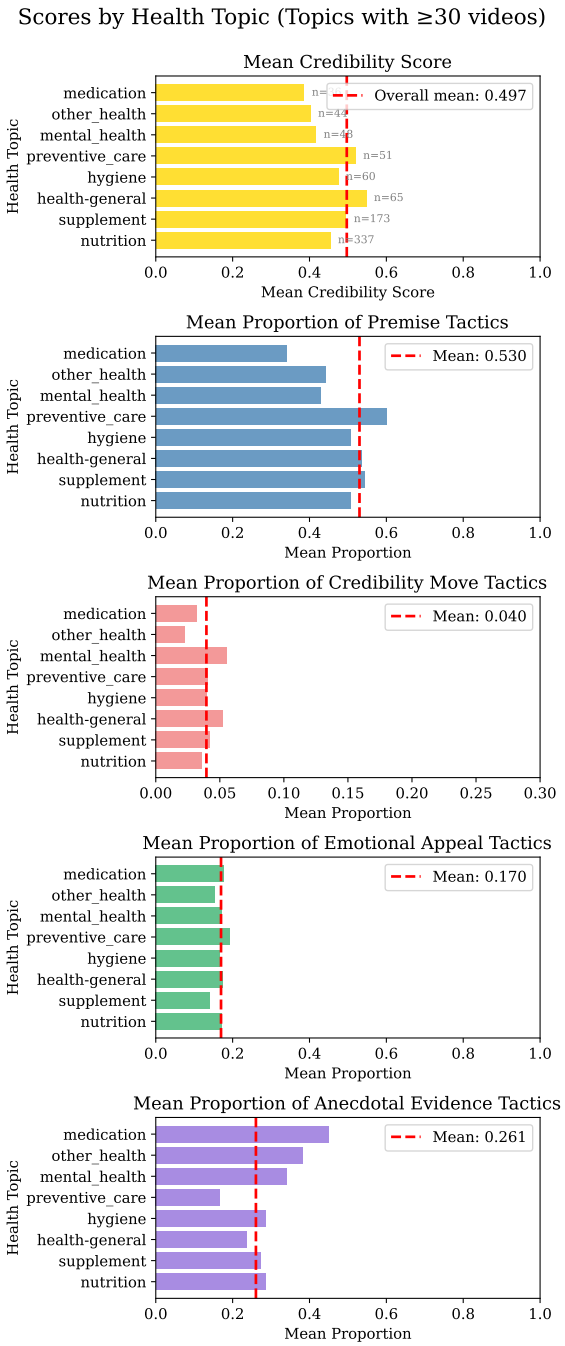

Figure 6: We show the top Health Topics and their various scores for the full set of 1,430 videos.

B Hyperparameters/Implementation Details

Section B.2 provides the hyperparameters and implementation details associated with each stage in our pipeline.

B.1 Dataset Curation.

We describe the detailed procedure used to construct the TikTok health influencer dataset summarized in Section 4. We adopt a hashtag-based data

collection strategy, where the hashtags are based on the ScienceFeedback website.

Specifically, for each ScienceFeedback article, we provide an LLM with the article text and the main *claim* under evaluation, and prompt it to generate up to five hashtags that are likely to be used by social media users when discussing the claim. These hashtags are intended to capture colloquial terminology commonly employed in influencer discourse.

Using the resulting hashtag set, we collect TikTok videos posted by creators with at least 1,000 followers, ensuring that the dataset focuses on influencer-generated content with non-trivial audience reach. Across all hashtags, we aggregate the retrieved videos and remove duplicates.

We obtain text transcriptions for each video using OpenAI's Whisper (Radford et al., 2023) API. To ensure topical relevance, we prompt an LLM to filter out videos that: (a) do not discuss health-related content, and (b) are not in English language. After this filtering step, the final dataset consists of 1,430 health-related TikTok videos whose transcriptions are in English.

The resulting corpus exhibits a diverse range of influencer popularity levels and communication styles. Notably, a substantial fraction of videos convey health advice implicitly rather than through explicit recommendations, motivating the takeaway-centric modeling approach proposed in this work.

Details regarding the construction of the annotated evaluation subset, including quality control, are provided in Appendix C.1.

B.2 TAIGR

Argumentation Structure. While constructing the argumentation graph, we treat the *edge weights* between different node pairs differently, ensuring the overall quality of resulting influencer reasoning chains for the *takeaway*. Specifically, we accept a candidate edge between a claim-node and the *takeaway*-node if its edge weight is greater than a threshold of 0.5, ensuring only strong directly supporting nodes to be associated with the *takeaway* while ignoring noise. This threshold is lowered to 0.25 when dealing with an edge between a pair of claim-nodes, accounting for those potentially implicit claims, which the influencer indirectly employs while arguing for the takeaway. For instance, a claim node: "this remedy eliminates worms" should be connected by an indirectly supporting claim node: "worms are harmful and are worth eliminating". Finally, we set a high threshold of 0.5 for a pair of statement nodes to ensure a high-quality reasoning chain while ignoring noise from the video transcriptions.

Trust Inference. Implementation details and hyperparameters for this stage of our pipeline are provided below.

PubMed data indexing. We index all PubMed articles from the last decade–2014 to 2025, for augmenting our argumentation graph with trustworthy scientific biomedical evidence. In total, our index consisted of $\approx 14M$ articles collected from PubMed annual baseline (2025) [7], and filtered based on: (a) *language*–restricted our index to English-only language; (b) *date*–restricted our index to articles from the last decade, ensuring it reflects latest medical guidelines and research.

Retrieval. For data indexing and retrieval, we utilized the codebase from RAGatouille [8], which internally uses ColBERTv2 (Santhanam et al., 2022) for data retrieval. After retrieval, we perform reranking using monot5-base-med-msmarco, where we employ a threshold of 0.3 (derived empirically by considering the rejection rate and relevance of the retrieved articles).

Factor-graph based inference. We represent each node $v \in V$ in the augmented argumentation graph as a random variable Y_v with 11 states $\{0, 0.1, 0.2, ..., 1.0\}$ with higher values indicating greater trustworthiness. We define unary and binary potentials to enforce global consistency in the graph structure and enable trust propagation from high-trust (PubMed) nodes to the rest of the nodes in the graph, including the *takeaway* node. Through unary potentials, we enforce PubMed evidence nodes to have a high-trust score $Y_v = 1$ by defining a hard constraint. Similarly, for the rhetorical node types {Anecdotal Evidence, Credibility Move, Emotional Appeal}, we enforce a low-trust score $Y_v = 0.2$.

To estimate the influence of support and attack potentials (binary potentials), we estimate β, γ variables using a small sample. Specifically, we randomly sample 20 datapoints, which follow the same distribution as the labeled dataset (70% from *incorrect* class, 15% from *partially correct* class, 15% from *correct* class); and use a grid-search to estimate the best-performing values for both variables β, γ. We

[7] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/

[8] https://github.com/AnswerDotAI/RAGatouille

used: $[0.1, 0.2, 0.3, ...1.0, 2.0, 3.0, ..., 10.0]$ grid and found that $\{\beta = 0.3, \gamma = 5\}$ achieved the highest F1-macro points.

For the final classification, we consider the assignment on the *takeaway* node. If the score is in the range $[0, 0.5]$, we classify the input transcription as *incorrect*. If it is in the range $[0.5, 0.7]$, we classify the transcription as *partially correct*. Beyond 0.7, we classify the transcription as *correct*.

Other Details. We use `GPT-4.1-mini` with $\{\text{temperature} = 0.25, \text{top_p} = 0.95\}$ for every LLM call in our pipeline.

C Experimental Setup

C.1 Constructing Annotated Dataset.

To evaluate the models on the *content validation* task, we construct a labeled dataset, using expert annotations from `ScienceFeedback` as a proxy. `ScienceFeedback` uses medical experts to validate real-world claims. Often times, it includes "viral" claims from online social media such as TikTok or YouTube. After validating a claim, medical experts provide a verdict, typically belonging to one of the pre-defined 18 classes.

We develop a labeled dataset by matching the main recommendation or *takeaway* from an influencer discourse with the annotated claim from `ScienceFeedback`. First, we embed all the annotated claims and set of all takeaways associated with each influencer transcription using `all-mpnet-base-v2`[9] (Reimers and Gurevych, 2019; Song et al., 2020). For each takeaway, we identify the most similar annotated claim based on cosine similarity, which is above a pre-defined threshold of 0.6. For the pairs that meet the thresholding criteria, we prompt an LLM to validate if the annotated claim entails the *takeaway*. Specifically, we use the prompt detailed in Figure E.2.1 for the classification. By doing this, we identified 185 potentially semantically equivalent {takeaway, annotated claim} pairs.

To ensure semantic equivalence and reliability, we perform human annotation over these pairs to discard the ones where the takeaway does not accurately correspond to the annotated claim, resulting in 166 high-quality pairs. The labels for the annotated claims in `ScienceFeedback` is generally skewed and is dominated by *Inaccurate/Unsupported* class, and the predefined 18 classes creates a class imbalance. Therefore, we collapse the 18 classes into 3 classes. Specifically, we map: (1) {*inaccurate, incorrect, flawed reasoning, unsupported, misleading*} classes from `Sciencefeedback` to ***incorrect*** class; (2) {*lacks context, imprecise, partially correct*} classes to ***partially correct*** class; and (3) {*mostly accurate, mostly correct, accurate, correct*} to ***correct*** class.

Consequently, among the 166 pairs, we identified 138 pairs belonged to *incorrect* class, 41 pairs belonged to *partially correct* class, and 16 pairs belonged to *correct* class. To increase the coverage of *correct* examples, we augment the dataset with 29 additional video transcripts sourced from `Cleveland Clinic`, a medically credible authority, which are all labeled as *correct*. Therefore, in total, our evaluation dataset consists of 195 pairs: 138 incorrect pairs, 41 partially correct pairs and 45 fully correct pairs. The influencer transcription gets the same label as that of the takeaway.

Aside, in this case, we note that the human annotator was a graduate STEM-student, who was not an author of the paper and was under the age of 30.

C.2 Baselines.

Given the healthcare domain of influencer transcripts, we avoid setups that rely on unconstrained end-to-end prompting without external evidence. Such approaches are known to hallucinate or rely on unverifiable reasoning (Lin et al., 2022). Instead, all baselines retrieve biomedical evidence from PubMed, ensuring that validation decisions are grounded in scientific literature.

Claim-centric RAG. This baseline decomposes a transcript into factual claims using an LLM, retrieves the top-30 PubMed articles for each claim, and classifies the retrieved evidence as supporting or opposing. The LLM then predicts the overall veracity of the transcript based on the aggregated evidence. This setup mirrors common claim-centric verification pipelines but does not explicitly model the central recommendation conveyed by the influencer.

RAG-w-Takeaway. This baseline introduces limited structure by first extracting the video's *takeaway* and retrieving the top-30 PubMed articles relevant to it. The retrieved evidence is classified as supporting or opposing, and the LLM is prompted to predict the veracity of the *takeaway* given this evidence. This baseline isolates the effect of takeaway-aware retrieval while omitting explicit modeling of claims and their interactions.

[9] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

LOKI. We adapt (Li et al., 2025) (open-sourced) as a claim-centric baseline, which aggregates the evidence from verifying each claim in order to verify the entire document. Our implementation is based on their official repository[10]. However, when retrieving evidence for a query, rather than using `SERPER-API`, we use our retrieval pipeline to obtain evidence from PubMed data, ensuring a fair comparison across all methods. Additionally, we note that (Li et al., 2025) originally used it for binary classification of a claim or a document. Specifically, (Li et al., 2025) used a threshold of 0.8, below which any claim was classified *incorrect*. We adapt this thresholding to adapt for the 3-class classification problem. By fixing the incorrect class threshold to 0.8, we empirically obtain a threshold for the *partially correct* class. Performing a grid-search over $[0.8, 0.81, 0.82, ..., 0.99]$, we found that a threshold of 0.95 achieves the highest F1-macro points, making it the threshold for *partially correct* class.

D Statistical Significance Test

Impact of takeaway type. We perform a paired bootstrap test ($10,000$ iterations) to compare our method against LOKI (the second-best model). For the implicit takeaway type, this resulted in a p-value of 0.022, indicating a statistically significant improvement.

However, for the explicit takeaway type, the test yielded a p-value of 0.136, indicating that the performance difference is not statistically significant. This indicates that while both models perform comparably on the classification task, our method offers a distinct advantage in scenarios requiring deeper inference regarding influencer's argumentation structure.

E Prompt Templates

E.1 Our Methodology

In this section, we provide all the prompt templates used in each stage of our pipeline: (1) Section E.1.1 shows the prompt templates used in *takeaway* identification stage; (2) Section E.1.2 shows the prompts used to model **how** the influencer argues for the takeaway; (3) Section E.1.3 shows the prompts used during the trust inference stage.

[10] https://github.com/Libr-AI/OpenFactVerification

E.1.1 *Takeaway*: What is the influencer trying to communicate?

Figure 7, 8 shows the prompts used for takeaway extraction and classification, respectively.

Prompt: *Takeaway* extraction.

```
### Task:
You are a text analyzer. From the given transcript, extract
      exactly **ONE key takeaway** that captures the most
      important recommendation, suggestion, protocol, or
      advice for the audience to remember.

### Definitions:
- *Key takeaway:* a single, complete piece of advice or
      instruction (explicit or implicit).
- *Explicit advice:* direct instruction (e.g., "Exercise 30
      minutes daily", "Take vitamin D").
- *Implicit advice:* indirect or suggestive phrasing that
      still nudges the audience toward action (e.g., "This
      worked for me", "Hope this helps everyone", "Studies
      suggest coffee reduces risk of X"). If it pushes the
      audience to act, treat it as a takeaway.

### Task Rules:
1. Extract exactly ONE *complete* key takeaway.
   - If multiple exist, choose the one most strongly directed
         at the audience.
   - If equally strong, choose the first one mentioned.
2. Ensure *contextual completeness*.
   - The **takeaway must be understandable in isolation** (no
         missing references like "this" or "that").
   - Rewrite slightly if needed for clarity, but keep meaning
         intact.
   - It should be complete and self-contained without any
         co-references.
3. Provide a justification.
   - Explain why this is the most important takeaway,
         referencing the transcript context.
4. Perform grounding.
   - Return the **exact substring(s)** from the transcript
         that contains the takeaway.
5. Fallback mechanism.
   - If no valid explicit or implicit takeaway is present,
         output the fallback JSON.

### Strict JSON Response Format:
```json
{{
  "takeaway": {{
    "text": "<insert-complete-takeaway-here>",
    "justification": "<why this takeaway is important>",
    "grounding": {{
      "text": "<exact-text-span(s)>"
    }}
  }}
}}
```
**If you cannot find a valid takeaway, respond with the
      following *fallback JSON structure*:**
```json
{{
  "takeaway": "NO_TAKEAWAY_FOUND"
}}
```

Input Transcript:
{transcript}
```

E.1.2 *Argumentation Structure*: How does the influencer support the takeaway?

Figures 9, 10, 11 shows the prompts used in the *preprocessing* phase. Figure 12 shows the prompt used in constructing the argumentation graph: identifying the potential support relationship between two nodes (between a claim-claim node, claim-takeaway node, and statement-statement node). A statement node can be one of the following {*Premise, Anecdotal Evidence, Credibility Move, Emotional Appeal*}.

Prompt: *Takeaway* classification.

```
### Task:
You are a text analyzer. You are also provided with a
      `takeaway`, which captures the most important
      recommendation, suggestion, protocol, or advice for the
      audience to remember from the provided `transcript`.
      Your task is to classify if the `takeaway` is
      "explicit" or "implicit" with respect to the
      `transcript`.

### Definitions:
- *Key takeaway:* a single, complete piece of advice or
      instruction (explicit or implicit).
- *Explicit advice:* direct instruction (e.g., "Exercise 30
      minutes daily", "Take vitamin D").
- *Implicit advice:* indirect or suggestive phrasing that
      still nudges the audience toward action (e.g., "This
      worked for me", "Hope this helps everyone", "Studies
      suggest coffee reduces risk of X"). If it pushes the
      audience to act, treat it as a takeaway.

### Task Rules:
1. Read the `transcript` and `takeaway` carefully. If the
      `takeaway` is due to **explicit advice**, classify it
      as "explicit". If the `takeaway` is due to **implicit
      advice**, classify it as "implicit".
2. Provide a justification.
   - Explain why this classification is correct by providing
         an explanation that is grounded to the `transcript`.

### Strict JSON Response Format:
```json
{{
    "label": <"explicit" or "implicit">,
    "justification": <why this takeaway is "explicit" or
          "implicit" by providing evidence from the
          transcript>,
}}
```

Input:
- **`Transcript`**: {transcript}

- **`Takeaway`**: {takeaway}
```

E.1.3 *Trust Inference*: Is the influencer's argument accurate?

Figures 13, 14, 15 shows the prompts for identification of checkworthy statements, expanding search queries, and classification of retrieved evidence, respectively.

E.2 Experimental Setup

In this section, we provide all the prompt templates used for constructing a labeled dataset, and developing baselines.

E.2.1 Constructing labeled dataset

Prompt: *Preprocessing*: Obtaining standalone statements from the video transcriptions.

```
### Task:
You are an expert in text analysis. Your task is to rewrite
      the provided `text` as a list of **completely
      standalone sentences**, where each sentence can be
      understood on its own while preserving the original
      tone and point of view.

### Strict Requirements:
1. **Sentence Splitting:** Split the `text` into individual
      sentences.
2. **Coreference Resolution:**
  - **Preserve Point of View:** Maintain the original
        first-person perspective ("I," "my," "me"). Do
        **not** replace these with "the speaker."
  - **Resolve Ambiguity:** Replace pronouns (e.g., "he," "it,"
        "they") or vague references with the correct noun
        phrase from the `text`, so that each sentence is
        **fully self-contained**.
3. **Grounding:** Do not add, remove, or invent any
      information. For each output sentence, strictly ensure
      that it is **fully grounded** in the `text`.
4. **Inclusion:** Every sentence in the original `text` must
      appear in the output.

### Output Format (Strict JSON):
```json
{{
  "all_standalone_sentences": [
    "<completely standalone sentence strictly grounded in the
          provided `text`>",
    ... # strictly include **all** sentences from the `text`,
          no information should be omitted
    ]
}}
```

Example:
`Input Text`: Okay, I know this is gonna sound wild —but
      your own urine might be one of the most powerful
      natural boosters out there. Think about it: it's made
      by you, packed with your own minerals, hormones, and
      amino acids —literally personalized medicine. I've been
      using aged urine topically for years, just rubbing a
      little on my skin before going out in the sun. I swear,
      the warmth activates it —my muscles feel fuller, my
      energy spikes, and my skin gets this crazy healthy
      glow. And yeah, sometimes I'll even take a sip of my
      morning urine. Sounds gross, but the energy and clarity
      I get after? Unreal. I'm almost 60, don't use sunscreen
      or steroids, and I still feel strong and alive. It's
      not for everyone, but for me, it's been a total
      game-changer. Sometimes the weirdest things nature
      gives us are the ones that actually work.

`Output JSON`:
```json
{{
  "all_standalone_sentences": [
    "Okay, I know this is gonna sound wild —but your own urine
          might be one of the most powerful natural boosters
          out there.",
    "Your urine is made by you and is packed with your own
          minerals, hormones, and amino acids —literally
          personalized medicine.",
    "I have been using aged urine topically for years by
          rubbing a little of the aged urine on my skin
          before going out in the sun.",
    "I swear that the warmth of the sun activates the aged
          urine.",
    "When the warmth activates the aged urine, my muscles feel
          fuller, my energy spikes, and my skin gets this
          crazy healthy glow.",
    "Sometimes I even take a sip of my morning urine.",
    "Drinking my morning urine sounds gross, but the energy
          and clarity I get after drinking it are unreal.",
    "I am almost 60 years old, do not use sunscreen or
          steroids, and I still feel strong and alive.",
    "Using your own urine is not for everyone, but for me,
          using it has been a total game-changer.",
    "Sometimes the weirdest things that nature gives us are
          the ones that actually work."
  ]
}}
```
**From the example output JSON, it is important to note that
      every sentence from the input text is included, every
      sentence is standalone, preserving sentence-level
      self-containment, and original tone and point of view.**

Input:
`text`: {transcript}
```

Prompt: *Preprocessing*: Classifying each statement based on its rhetorical functionality.

```
### Task:
You are an expert in argumentation theory and discourse analysis. Your goal is to classify each provided `Statement` into **exactly one** of
       the following categories, based on its rhetorical function and style of persuasion. Understand the provided `Transcript` to help you
       classify the statements.

---

### Classification Categories:

#### 1. "Premise" (Logos)
- **Definition:** A logical or factual statement that provides *reason-based* support for a claim.
- **Key Features:**
  - Uses reasoning, causation, or factual explanation ("because", "therefore", "since").
  - Does **not** rely on personal experience, emotion, or authority.
- **Example:** "Because urine contains urea and minerals, it may have antimicrobial properties."
- **Not Premise if:** The statement talks about personal experience, credentials, or emotional outcomes.

---

#### 2. "Anecdotal Evidence"
- **Definition:** A personal story, observation, or example based on individual experience rather than systematic proof.
- **Key Features:**
  - Mentions the speaker's or someone's *own experience* ("I tried", "my friend noticed", "people say").
  - Often used to illustrate or reinforce a point with a story.
- **Example:** "I started drinking urine last year, and my skin has never been clearer."
- **Not Anecdotal if:** The statement generalizes or reasons logically without a personal element.

---

#### 3. "Credibility Move" (Ethos)
- **Definition:** A statement that builds trust, expertise, or authority of the speaker to make the argument more convincing.
- **Key Features:**
  - References credentials, background, authority, or expertise ("As a doctor...", "I've studied this for years", "According to this
        research...").
  - Sometimes includes name-dropping or citing authoritative sources.
- **Example:** "As a practitioner of Ayurvedic medicine, I can assure you this technique is safe."
- **Not Credibility Move if:** It focuses on data, reasoning, or emotions rather than authority.

---

#### 4. "Emotional Appeal" (Pathos)
- **Definition:** A statement that evokes emotion (hope, fear, pride, shame, love, empowerment) to persuade.
- **Key Features:**
  - Uses emotionally charged language or imagery.
  - Focuses on how the audience *feels* rather than what is *true*.
- **Example:** "It's time to reclaim control of your body —don't let big pharma tell you what's natural."
- **Not Emotional Appeal if:** The argument is calm, factual, or authority-based.

---

#### 5. "None" (None)
- **Definition:** The statement doesn't clearly fit any of the categories above.
- **Examples:**
  - Off-topic comments or factual statements unrelated to persuasion.
  - Ambiguous fragments without clear function.

---

### Additional Guidelines:
1. **Prioritize rhetorical intent** over surface wording —ask: *What is the statement trying to do?*
2. **Avoid defaulting to "Premise"** unless the statement clearly relies on reasoning.
3. **When authority-based —prefer "Credibility Move".**
4. **When personal or experiential —prefer "Anecdotal Evidence".**
5. **When emotionally loaded —prefer "Emotional Appeal".**
6. **When unclear —use "None".**

---

### Strict JSON Output Format:
```json
{{
    "<statement_id>": {{ # `statement_id` should match the `id` of the statement in the given `statements` dict.
        "type": "<classification type>",
        "explanation": "<clear justification referring to its features>"
    }},
    ...
}}
```

Input:
`**Transcript**`: {full_transcript}

`**Statements**`: {statements_dict}
```

Prompt: *Preprocessing*: Claim identification – accounting for potentially implicit reasoning chains.

```
### Task:
You are an expert in structured argumentation analysis. Your goal is to identify **ALL the claims** being made in the given `transcript`.
      You are also provided with individual `statements` from the same `transcript`, each with a unique `id`.

### Definition of a Claim:
A **claim** is a statement that asserts a position, viewpoint, or conclusion that can be evaluated as true, false, or debatable. Claims are
      the central points of an argument. There are two types of claims:

- **Explicit Claim:** A statement directly expressed in the text where the speaker clearly asserts a position or conclusion.
- **Implicit Claim:** A statement that suggests, implies, or presupposes a position without directly stating it; the claim must be inferred
      from context, assumptions, causal reasoning, or the relationship between statements.

### Instructions:
- Read and understand the `transcript` and the `statements` carefully.
- Identify **both explicit and implicit claims**, including claims that describe causes, effects, assumptions, goals, motivations, or
      implied problems and solutions.
- For each statement in the transcript:
    - Decide if it expresses (explicitly or implicitly) a claim.
    - If it does, extract it and add it to the dictionary of identified claims.
- If multiple statements together express one single claim, combine them. Add the combined statement to the identified dictionary of claims.
- If a statement(s) is not making a claim, skip it.
- Ensure each identified claim is **different** from the others. We do not want the same claims being identified multiple times.
- When inferring **implicit claims**, look for:
    - Assumptions that make explicit claims meaningful (e.g., "worms are harmful" implied by "this remedy eliminates worms"; OR "urine is
          beneficial for health" implied by "applying urine to the skin improves skin health").
    - Implied causal or motivational links (e.g., "eliminating worms improves health"; OR "the warmth from sunlight activates beneficial
          properties in urine" implied by "the warmth activates it - muscles feel fuller and energy spikes").
    - Contextual premises or problem statements that justify a proposed action (e.g., "natural bodily substances can provide health benefits
          when reused correctly" implied by "your own urine might be one of the most powerful natural boosters out there").
- Each identified claim must be **self-contained** and understandable on its own, without referring back to the transcript.
- For every identified claim, record the following information to remain faithful to the given `transcript`:
    - `claim_id`: The <id> of the claim. Start from 1 and increment for each new claim. Format: `claim_<id>`.
    - `claim_text`: The text of the claim that can be understood on its own without referring to the `transcript`.
    - `claim_type`: The type of the claim ("explicit" or "implicit").
    - `statement_ids`: Include **all** the sentence ids that are part of the identified claim. **Ensure that the `ids` map to the ids of the
          statements in the `statements` dictionary.** This is important to ensure that the claim is faithful to the transcript.
    - `justification`: Provide a brief justification for why you identified this as a claim and whether it was explicitly stated or
          implicitly inferred.
    - `grounding_to_statements`: Provide a brief justification for how this claim is grounded to the `statement_ids` in the given
          `statements` dictionary, including reasoning about inferred or combined meaning if applicable.
- **If you cannot identify even a single claim, then return an empty dictionary.**

### Example showing extracted `claims` from a `transcript`:
**`Transcript`**:
Okay, I know this is gonna sound wild —but your own urine might be one of the most powerful natural boosters out there. Think about it: it's
      made by you, packed with your own minerals, hormones, and amino acids —literally personalized medicine. I've been using aged urine
      topically for years, just rubbing a little on my skin before going out in the sun. I swear, the warmth activates it —my muscles feel
      fuller, my energy spikes, and my skin gets this crazy healthy glow. And yeah, sometimes I'll even take a sip of my morning urine.
      Sounds gross, but the energy and clarity I get after? Unreal. I'm almost 60, don't use sunscreen or steroids, and I still feel strong
      and alive. It's not for everyone, but for me, it's been a total game-changer. Sometimes the weirdest things nature gives us are the
      ones that actually work.

**`Extracted Claims` (showing *only* the `claim_text` and `claim_type`)**:
1. Human urine is a powerful natural booster. (explicit)
2. Urine contains beneficial concentrations of minerals, hormones, and amino acids. (explicit)
3. Urine functions as personalized medicine because it is composed of an individual's own unique minerals, hormones, and amino acids.
      (implicit)
4. The warmth from sunlight activates the properties of topically-applied urine. (explicit)
5. Applying sun-activated, aged urine to the skin causes muscles to feel fuller, energy to spike, and skin to develop a healthy glow.
      (explicit)
6. Ingesting one's own morning urine results in increased energy and mental clarity. (implicit)
7. Natural bodily substances can have health benefits when reused correctly. (implicit)

It is important to note that each `claim_text` can be understood on its own without referring to the `transcript`. It does **not** contain
      any information such as "the speaker" or "I" or "my" or "me" or any other pronouns.

### Strict JSON Output Format:
```json
{{
    "<claim_1>": {{
        "claim_text": "Human urine is a powerful natural booster.",
        "claim_type": "explicit",
        "statement_ids": [S1],
        "justification": "The speaker asserts a specific viewpoint that human urine is a powerful natural booster. It is a debatable
              assertion that can be evaluated as true or false.",
        "grounding_to_statements": "S1 consists of the statement 'Your own urine might be one of the most powerful natural boosters out
              there.' which is a clear assertion that human urine is a powerful natural booster."
    }},
    ...
}}
```
If you cannot identify even a single claim, then return the following empty dictionary.
```json
{{}}
```

Input:
`Transcript`: {transcript}

`Statements`: {statements_dict}
```

Prompt: Building Argumentation Graph: Identifying relationship between two nodes.

```
### Task:
You are an expert argument analyst. You are provided with a
      `Claim to Analyze` and a list of `Available Claims`.
      For each claim in the list, your goal is to classify it
      into one of the following categories based on how well
      it supports the `Claim to Analyze`. The categories are:
      **direct support**, **weak support**, or **no
      support**, each of which is defined below. You are also
      provided with the `Transcript`, which you can use to
      get additional context about the `Claim to Analyze` and
      the claims in the list.

If a claim fits more than one category, assign it to the
      **strongest** one (for example, prefer *direct support*
      over *weak support* if both could apply; prefer *weak
      support* over *no support* if both could apply;).

### Definitions:
- **direct support:** Claim A *strongly directly supports*
      Claim B if A clearly *restates, proves, or provides
      explicit causal justification* for B, leaving little
      interpretive ambiguity.
   - Example: `Claim to Analyze` is "Drinking green tea helps
         with weight loss" and a directly supporting claim
         is "I lost weight after switching to green tea." or
         "Green tea boosts metabolism, which helps the body
         burn fat faster." (explicit causal justification)
- **weak support:** Claim A *weakly supports* Claim B if A is
      *topically relevant and consistent in stance*, but does
      not directly assert, prove, or justify B. The support
      is inferential or anecdotal rather than explicit.
   - This includes claims that highlight the importance,
         relevance, or consequences of the topic discussed
         in B. For example, stating that the issue requires
         attention, action, treatment etc. can increase the
         plausibility of B without proving it.
   - Example: `Claim to Analyze` is "Drinking green tea helps
         with weight loss" and an weakly supporting claim is
         "Lots of people drink it to feel lighter." or "It's
         part of many healthy diets."
   - Another example: `Claim to Analyze` is "Green tea boosts
         metabolism and burns fat." and an indirectly
         supporting claim is "Green tea supports gut health
         and digestion." (*causal link between digestion and
         fat-burning is implied, not stated*.)
- **no support:** Claim A *does not support* Claim B if there
      is *no meaningful logical, causal, or thematic
      connection* that would make B more believable,
      plausible, or credible.
   - Example: `Claim to Analyze` is "Drinking green tea helps
         with weight loss" and a *no support* claim is "I
         prefer coffee in the morning."

### For Each Selected Claim:
For every claim you select, provide:
- `"id"` —the claim ID from the available list. It must
      **strictly** be one of the IDs from the available list.
- `"category"` —the category of the claim ("direct support",
      "weak support", or "no support").
- `"explanation"` —why this claim supports/does not support
      the `Claim to Analyze` based on the definitions above.
      The explanation should be complete, concise, and clear.

### Strict JSON Output Format:
If you find supporting claims, return a **valid JSON** with
      the following format:
```json
{{
    "<claim-id>": {{
        "category": "<category of the claim>",
        "explanation": "<why this claim supports/does not
              support the `Claim to Analyze`>",
    }}
}}
```

Input:
`**Transcript**`: {full_transcript}
`**Claim to Analyze**`: {claim_text}
`**Available Claims JSON**`: {available_claims_json}
```

Prompt: Classifying whether a node is PubMed-checkworthy or not.

```
### Task:
You are an expert at understanding the provided text. You are
      provided with a `context` (transcript) and a several
      texts derived from the `context`. Your task is to
      classify each provided `text` as either
      `PubMed-checkworthy` or `not PubMed-checkworthy`.

### Decision Guide:
A text is **PubMed-checkworthy** if:
- It makes a **biomedical or health-related factual claim**
      that can be evaluated with scientific evidence (e.g.,
      about diseases, treatments, nutrients, physiology,
      hormones, biological effects, and many more).
- The claim is **generalizable beyond the specific provided
      context** and could, in principle, be **verified or
      refuted** using PubMed or biomedical literature.

A text is **not PubMed-checkworthy** if:
- It is **personal, anecdotal, or rhetorical** (e.g., "I love
      how it makes me feel").
- It is **too vague or context-specific** to be scientifically
      verified (e.g., "this is aged urine").
- It is **non-biomedical** (e.g., about opinions, beliefs,
      emotions, aesthetics, or general lifestyle).

### Classification Rules:
1. Analyze the provided `context` to get a sense of the
      overall context associated with each `text`.
2. For each `text`, decide if it expresses a verifiable
      biomedical claim.
3. When classifying a `text`,
   - if multiple labels are applicable, prioritize
         `PubMed-checkworthy` over `not PubMed-checkworthy`.
   - if in doubt, classify as `PubMed-checkworthy`.
4. Output one JSON entry per text, preserving its exact `id`.
      This `id` must strictly match the `id` of the `text` in
      the input.
5. Strictly ensure that each `id` provided in the input is
      present in the output.

### Strict JSON Output Format:
```json
{{
  "<text_id>": {{
    "label": "PubMed-checkworthy" | "not PubMed-checkworthy",
    "explanation": "<brief reason justifying the
          classification decision>"
  }}
}}
```

Example:
`Transcript`: "This is my aged urine. Okay, I know this is
      gonna sound wild —but your own urine might be one of
      the most powerful natural boosters out there. Think
      about it: it's made by you, packed with your own
      minerals, hormones, and amino acids —literally
      personalized medicine. I've been using aged urine
      topically for years, just rubbing a little on my skin
      before going out in the sun. I swear, the warmth
      activates it —my muscles feel fuller, my energy spikes,
      and my skin gets this crazy healthy glow. And yeah,
      sometimes I'll even take a sip of my morning urine.
      Sounds gross, but the energy and clarity I get after?
      Unreal. I'm almost 60, don't use sunscreen or steroids,
      and I still feel strong and alive. It's not for
      everyone, but for me, it's been a total game-changer.
      Sometimes the weirdest things nature gives us are the
      ones that actually work."

`Texts to classify`: {{
   "claim_1": "Aged urine boosts testosterone and improves
         skin health.",
   "S2": "This is aged urine."
}}

`Output`: {{"claim_1": {{
   "label": "PubMed-checkworthy",
   "explanation": "This is a biomedical claim about hormonal
         and skin effects that can be verified using
         scientific literature."
}},
"S2": {{
   "label": "not PubMed-checkworthy",
   "explanation": "This is a descriptive statement tied
         specifically to the provided context, not a
         biomedical claim that can be checked against
         PubMed."
}}}}

Input:
`Transcript`: {transcript}
`Texts to classify`: {nodes_json}
```

Prompt: Expanding search queries for checkworthy nodes.

```
You are an expert at generating short *semantic* search queries for biomedical literature retrieval (for use with embedding-based / semantic
      search systems). Given a `transcript` (context) and a `claim` already marked PubMed-checkworthy, generate concise, natural-language
      queries that maximize semantic recall for both **supporting** and **opposing** evidence.

### Output goals:
- Produce up to **5 supporting** and up to **5 opposing** semantic queries.
- Queries must be **natural-language**, standalone, and optimized for embedding-based retrieval (no boolean operators, no Medline field
      tags).
- Each query should be paired with a one-sentence **justification** explaining what kind of evidence it targets.

### Query construction heuristics (semantic best-practices):
1. **Natural phrasing**: write queries as short sentences or questions a researcher might ask (e.g., "Topical human urine effect on skin
      hydration in adults?"). Avoid using AND/OR, quotes, field tags, or forced Boolean structure.
2. **Diverse angles**: for both supporting and opposing sets, vary the focus across:
   - Outcome/effect (e.g., testosterone, skin appearance, muscle recovery)
   - Mechanism/biomarkers (e.g., hormones, amino acids, immune markers)
   - Population and application route (topical vs oral; adults vs older adults)
   - Study design/strength (randomized controlled trial, clinical trial, observational study, review/meta-analysis)
   - Safety/adverse events and null findings (no effect, adverse outcomes)
3. **Paraphrase strategy**: include different grammatical forms —short declarative, question form, and mechanism-focused phrase. This
      increases embedding-space coverage.
4. **Contrast language for opposing queries**: use explicit negation or null-effect language (e.g., "no increase", "no association",
      "ineffective") and terms like "lack of benefit", "no effect", "adverse" to surface contradictory or null-result papers.
5. **Conciseness**: aim for approximately 5-12 words per query (short enough to get focused embeddings, long enough to include key concepts).
6. **No repetition**: each query must add a new angle or emphasize a different keyword set (in natural language), not just reword the same
      phrase.
7. **Scope alignment**: all generated queries must stay *strictly within the scope of the provided claim*. Do not introduce new effects,
      outcomes, mechanisms, or populations not mentioned or implied by the claim. For example, if the claim concerns "Topical aged urine
      increasing testosterone and improving skin appearance", queries about "reducing brain fog" or unrelated outcomes are invalid.

### Strict Output JSON format:
Return valid JSON:
```json
{{
  "supporting_queries": [
    {{"query_id":"S1","query_text":"<natural-language semantic query>","justification":"<why this finds supporting evidence, and why is the
          scope aligned with the claim>"}},
    ...
  ],
  "opposing_queries": [
    {{"query_id":"O1","query_text":"<natural-language semantic query>","justification":"<why this finds opposing/null evidence, and why is
          the scope aligned with the claim>"}},
    ...
  ]
}}
```
If you are unable to generate any supporting queries, return an empty list for the "supporting_queries" key. Similarly, if you are unable to
      generate any opposing queries, return an empty list for the "opposing_queries" key. If you are unable to generate any queries, return
      an empty list for both the "supporting_queries" and "opposing_queries" keys.

Example:
Claim: "Topical aged urine increases testosterone and improves skin appearance."

Supporting queries (examples):
```json
{{"query_id":"S1", "query_text":"Topical urine application and testosterone changes in adults", "justification":"Targets studies measuring
      hormone changes after topical urine exposure. It is specific to the provided claim. The scope of the query is within the context of
      the provided claim."}}
{{"query_id":"S2","query_text":"Effects of topical application of human urine on skin hydration and color","justification":"Catches
      dermatology studies on topical biological secretions and skin outcomes.This is within the context of the provided claim."}}
```
Opposing queries (examples):
```json
{{"query_id":"O1","query_text":"Clinical trials showing no testosterone increase after topical application of human
      urine","justification":"Finds RCTs or trials reporting no hormonal effects from topical interventions. It is specific to the provided
      claim. The scope of the query is within the context of the provided claim."}}
{{"query_id":"O2","query_text":"Adverse outcomes or infection risk from applying human bodily fluids such as urine to
      skin","justification":"Targets safety and adverse-event literature about topical body-fluid use. It is specific to the provided claim.
      The scope of the query is within the context of the provided claim."}}
```

Input:
`Transcript`: {transcript}
`Claim`: {claim}
```

Prompt: Classification of retrieved evidence into: {*strong support, weak support, neutral, weak oppose, strong oppose*}.

```
### Task:
You are an expert at natural language inference and scientific stance classification. You are provided with a `Claim` and a list of
      potentially relevant `evidence articles`. For each evidence in the list, your goal is to classify it into one of the following
      categories, reflecting how strongly the evidence relates to and aligns with the `Claim`:
**strong support**, **weak support**, **neutral**, **weak oppose**, or **strong oppose**.
You are also provided with the `Transcript`, which gives broader context about the `Claim` and may help you classify the evidence.

### Definitions:
- **strong support:**
   - Evidence is *directly consistent* with the claim and provides **explicit empirical or causal confirmation** of the same relationship,
         typically through high-quality experimental, clinical, or meta-analytic findings that show a statistically significant and
         meaningful effect.
   - Such evidence *affirms* the claim's key variables, directionality, and population with high certainty.
   - *Example:*
      - Claim: "Green tea reduces blood pressure in adults."
      - Evidence: "A large randomized controlled trial found that daily green tea consumption significantly decreased both systolic and
            diastolic blood pressure in adults."
      - Explanation: **strong support** because of explicit causal agreement from a high-quality study design, affirming the key variables
            and population.
- **weak support:**
   - Evidence is *broadly consistent* with the claim but provides **indirect, correlational, or partial justification**, rather than direct
         and complete proof.
   - This includes mechanistic reasoning, findings from different but related populations (e.g., animal models for a human claim), or
         evidence that confirms only a part of the claim.
   - *Example 1 (Indirect Population):*
      - Claim: "Green tea reduces blood pressure in adults."
      - Evidence: "In a study on hypertensive rats, consumption of green tea extract led to a significant reduction in blood pressure."
      - Explanation: **weak support** as the evidence is from an animal model, offering only indirect, inferential support for the human
            claim.
   - *Example 2 (Partial Confirmation):*
      - Claim: "Green tea reduces blood pressure in adults."
      - Evidence: "A clinical study observed that green tea consumption was associated with a statistically significant decrease in
            systolic blood pressure, but not diastolic blood pressure."
      - Explanation: **weak support** because the evidence only partially confirms the claim's outcome (one measure of blood pressure, but
            not the other).
- **neutral:**
   - Evidence provides **no clear stance** toward the claim's truth, either because it is *topically unrelated* or discusses a *logically
         irrelevant* aspect of the topic.
   - It neither supports nor contradicts the claim. Any evidence that is not relevant to the claim's central assertion must be classified as
         neutral.
   - *Example 1 (Topically Related but Irrelevant):*
      - Claim: "Green tea reduces blood pressure in adults."
      - Evidence: "Green tea enhances cognitive alertness."
      - Explanation: **neutral** as the evidence shares the topic but is irrelevant to the claimed relationship with blood pressure.
   - *Example 2 (Topically Unrelated):*
      - Claim: "Green tea reduces blood pressure in adults."
      - Evidence: "Regular cardiovascular exercise is effective for improving heart health."
      - Explanation: **neutral** because the evidence is on a completely different topic and has no bearing on the claim about green tea.
- **weak oppose:**
   - Evidence is *topically relevant* but presents **inconclusive, mixed, or weakly contradictory findings**.
   - It may report a lack of a statistically significant effect from a study that may be underpowered, or present findings with limitations
         that cast doubt without conclusively refuting the claim.
   - *Example:*
      - Claim: "Green tea reduces blood pressure in adults."
      - Evidence: "A small pilot study with 30 participants found no statistically significant change in blood pressure after four weeks of
            green tea supplementation."
      - Explanation: **weak oppose** because the findings from a small, potentially underpowered study fail to support the claim but do not
            have enough statistical power to strongly refute it.
- **strong oppose:**
   - Evidence *directly contradicts* the claim by providing **robust empirical refutation or inverse findings**.
   - This includes conclusive null effects from large, high-powered studies or meta-analyses, or significant findings showing an effect in
         the opposite direction.
   - *Example:*
      - Claim: "Green tea reduces blood pressure in adults."
      - Evidence: "A meta-analysis of 15 randomized controlled trials, including over 2,000 participants, found no statistically
            significant effect of green tea consumption on blood pressure."
      - Explanation: **strong oppose** because a large-scale, high-quality evidence synthesis explicitly and conclusively finds a null
            effect, refuting the claim.

### For Each Evidence:
For every evidence, provide:
- `"id"` —the evidence ID from the available list. It must **strictly** match one of the provided IDs.
- `"category"` —one of the five stance labels (`"strong support"`, `"weak support"`, `"neutral"`, `"weak oppose"`, or `"strong oppose"`).
- `"explanation"` —a concise justification explaining *why* this evidence fits that stance, referencing causal direction, strength of
      association, and topical relevance. Your explanation should be precise, especially for boundary cases like partial support or indirect
      evidence from different populations.

### Strict JSON Output Format:
Return valid JSON in the following structure:
```json
{{
   "<evidence-id>": {{
      "category": "<strong support | weak support | neutral | weak oppose | strong oppose>",
      "explanation": "<why this evidence fits this category>"
   }}
}}
```

Input:
`**Transcript**`: {transcript}
`**Claim**`: {claim_text}
`**Available Evidence**`: {evidence_formatted}
```

Prompt: Entailment between {*Takeaway*-annotated claim} for constructing labeled dataset.

```
You are an entailment expert. Given `Text-1` and `Text-2`,
      decide if `Text-2` entails `Text-1`.

### Key Rule:
- Ignore any recommendations, advice, instructions, or
      protocols in the texts.
- Focus only on the **core factual content** (claims about
      properties, causes, or effects).
- `Text-2` entails `Text-1` if everything asserted in
      `Text-1` is logically implied by `Text-2`.

---

### Examples

#### Example 1:
- Text-1: "Avoid wearing AirPods close to your brain because
      Bluetooth radiation can cause health issues."
- Text-2: "AirPods emit dangerous levels of radiation that
      harm the body."
- After removing the advice from Text-1, the core is:
      "Bluetooth radiation from AirPods can cause health
      issues."
- Text-2 asserts that AirPods emit harmful radiation. This
      logically implies the core of Text-1.
- **Answer:** entails = yes.

---

#### Example 2:
- Text-1: "To support your immune system and help prevent or
      treat infections during sick season, make a natural
      remedy called 'Nature's Amoxicillin' by blending onion,
      ginger, turmeric, garlic, lemon with peel, honey,
      cayenne, and other potent antiviral and antibacterial
      ingredients, and consume it as a wellness shot."
- Text-2: "A soup made of garlic, onions, thyme, and lemon can
      replace the flu shot and cure other illnesses such as
      the common cold and norovirus."
- Removing the advice from Text-1, the core is: "A mixture of
      certain natural ingredients (onion, ginger, turmeric,
      garlic, lemon, etc.) can help prevent or treat
      infections during sick season."
- Text-2 asserts that a mixture of similar ingredients
      (garlic, onions, thyme, lemon) can cure illnesses like
      the common cold.
- Both describe natural mixtures of overlapping ingredients
      with the shared claim that these mixtures prevent or
      cure infections.
- **Answer:** entails = yes.

---

#### Example 3:
- Text-1: "Taking one tablespoon daily of the prepared mixture
      containing ginger, red onion, garlic, turmeric, black
      pepper, lemon or lime, and honey will boost your immune
      system by 200%, reduce body inflammation, and protect
      you from a long list of diseases."
- Text-2: "A soup made of garlic, onions, thyme, and lemon can
      replace the flu shot and cure other illnesses such as
      the common cold and norovirus."
- Removing the advice from Text-1, the core is: "A natural
      mixture with garlic, onion, ginger, turmeric, etc.,
      protects against many diseases and strengthens the
      immune system."
- Text-2 asserts that a natural soup with overlapping
      ingredients (garlic, onion, lemon) cures illnesses and
      can replace the flu shot.
- Both describe natural mixtures with protective health
      effects, so Text-2 logically implies the general claim
      in Text-1.
- **Answer:** entails = yes.

---

### Output Format (strict JSON):
```json
{{
  "entails": "<yes or no>",
  "justification": "<short explanation>",
  "evidence": "<if entails=yes, quote the supporting part of
        Text-2; else leave empty>"
}}
```

Input:
`Text-1`: {text_1}
`Text-2`: {text_2}
```